CHAPTER FIVE

Energy Minimisation and Related Methods for Exploring the Energy Surface

5.1 Introduction

For all except the very simplest systems the potential energy is a complicated, multi-dimensional function of the coordinates. For example, the energy of a conformation of ethane is a function of the 18 internal coordinates or 24 Cartesian coordinates that are required to completely specify the structure. As we discussed in Section 1.3, the way in which the energy varies with the coordinates is usually referred to as the *potential energy surface* (sometimes called the *hypersurface*). In the interests of brevity all references to 'energy' should be taken to mean 'potential energy' for the rest of this chapter, except where explicitly stated otherwise. For a system with N atoms the energy is thus a function of $3N - 6$ internal or $3N$ Cartesian coordinates. It is therefore impossible to visualise the entire energy surface except for some simple cases where the energy is a function of just one or two coordinates. For example, the van der Waals energy of two argon atoms (as might be modelled using the Lennard-Jones potential function) depends upon just one coordinate: the interatomic distance. Sometimes we may wish to visualise just a part of the energy surface. For example, suppose we take an extended conformation of pentane and rotate the two central carbon–carbon bonds so that the torsion angles vary from 0° to 360°, calculating the energy of each structure generated. The energy in this case is a function of just two variables and can be plotted as a contour diagram or as an isometric plot, as shown in Figure 5.1.

We will use the term 'energy surface' to refer not only to systems in which the bonding remains unchanged, as in these two examples, but also where bonds are broken and/or formed. Our discussion will be appropriate to both quantum mechanics and molecular mechanics, except where otherwise stated.

In molecular modelling we are especially interested in minimum points on the energy surface. Minimum energy arrangements of the atoms correspond to stable states of the system; any movement away from a minimum gives a configuration with a higher energy. There may be a very large number of minima on the energy surface. The minimum with the very lowest energy is known as the *global energy minimum*. To identify those geometries of

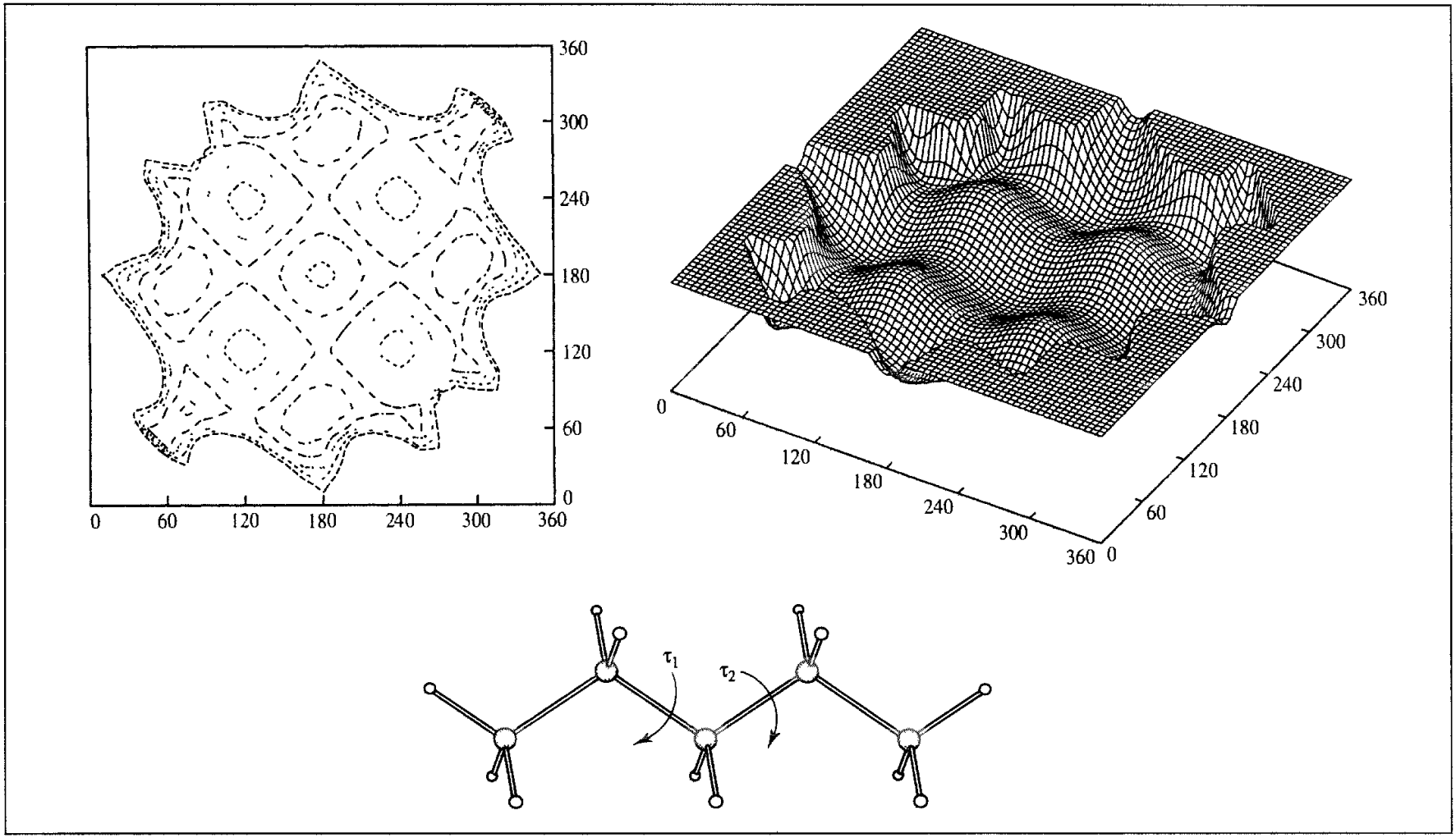

Fig. 5.1 Variation in the energy of pentane with the two torsion angles indicated and represented as a contour diagram and isometric plot Only the lowest-energy regions are shown.

the system that correspond to minimum points on the energy surface we use a *minimisation algorithm*. There is a vast literature on such methods and so we will concentrate on those approaches that are most commonly used in molecular modelling. We may also be interested to know how the system changes from one minimum energy structure to another. For example, how do the relative positions of the atoms vary during a reaction? What structural changes occur as a molecule changes its conformation? The highest point on the pathway between two minima is of especial interest and is known as the *saddle point*, with the arrangement of the atoms being the *transition structure*. Both minima and saddle points are stationary points on the energy surface, where the first derivative of the energy function is zero with respect to all the coordinates.

A geographical analogy can be a helpful way to illustrate many of the concepts we shall encounter in this chapter. In this analogy minimum points correspond to the bottom of valleys. A minimum may be described as being in a 'long and narrow valley' or 'a flat and featureless plain'. Saddle points correspond to mountain passes. We refer to algorithms taking steps 'uphill' or 'downhill'.

5.1.1 Energy Minimisation: Statement of the Problem

The minimisation problem can be formally stated as follows: given a function f which depends on one or more independent variables $x_1, x_2, \ldots, x_i$, find the values of those variables where f has a minimum value. At a minimum point the first derivative of the function with respect to each of the variables is zero and the second derivatives are all positive:

$$\frac{\partial f}{\partial x_i} = 0; \qquad \frac{\partial^2 f}{\partial x_i^2} > 0 \tag{5.1}$$

The functions of most interest to us will be the quantum mechanics or molecular mechanics energy with the variables x_i being the Cartesian or the internal coordinates of the atoms. Molecular mechanics minimisations are nearly always performed in Cartesian coordinates, where the energy is a function of $3N$ variables; it is more common to use internal coordinates (as defined in the Z-matrix) with quantum mechanics. For analytical functions, the minimum of a function can be found using standard calculus methods. However, this is not generally possible for molecular systems due to the complicated way in which the energy varies with the coordinates. Rather, minima are located using numerical methods, which gradually change the coordinates to produce configurations with lower and lower energies until the minimum is reached. To illustrate how the various minimisation algorithms operate, we shall consider a simple function of two variables: $f(x, y) = x^2 + 2y^2$. This function is represented as a contour diagram in Figure 5.2. The function has one minimum point, located at the origin. In our examples we will attempt to locate the minimum from the point (9.0, 9.0). Although this is a function of just two variables for the purposes of illustration, all of the methods that we shall consider can be applied to functions of many more variables.

We can classify minimisation algorithms into two groups: those which use derivatives of the energy with respect to the coordinates and those which do not. Derivatives can be useful

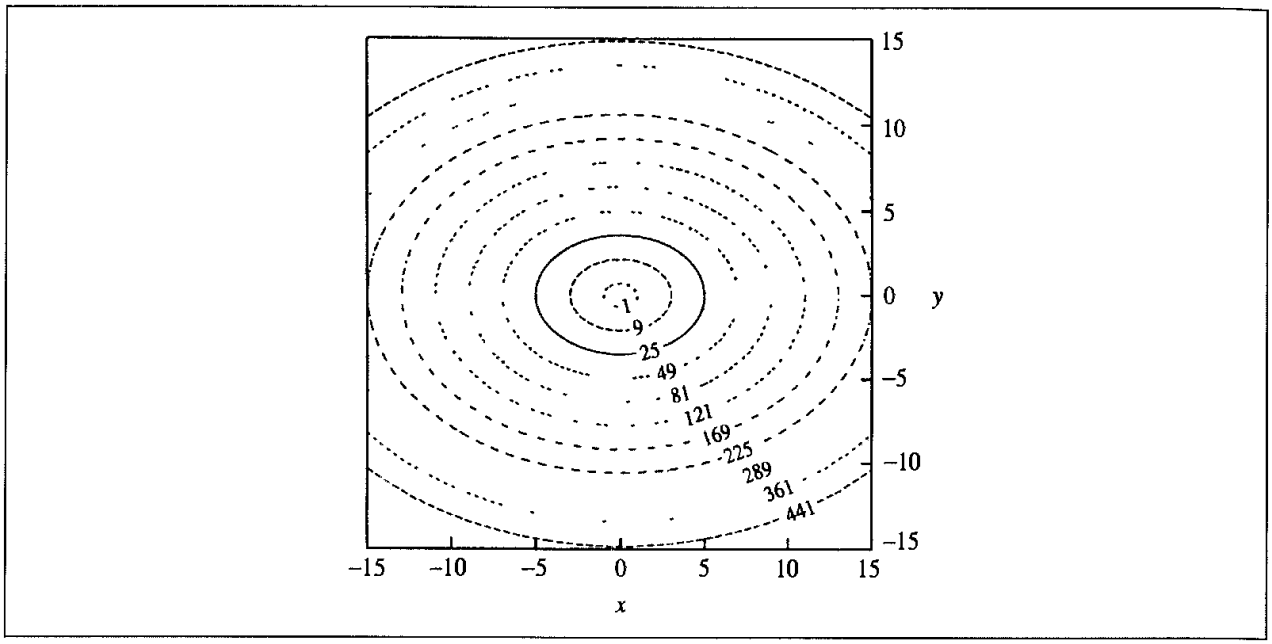

Fig 5 2· The function $x^2 + 2y^2$.

because they provide information about the shape of the energy surface, and, if used properly, they can significantly enhance the efficiency with which the minimum is located. There are many factors that must be taken into account when choosing the most appropriate algorithm (or combination of algorithms) for a given problem; the ideal minimisation algorithm is the one that provides the answer as quickly as possible, using the least amount of memory. No single minimisation method has yet proved to be the best for all molecular modelling problems and so most software packages offer a choice of methods. In particular, a method that works well with quantum mechanics may not be the most suitable for use with molecular mechanics. This is partly because quantum mechanics is usually used to model systems with fewer atoms than molecular mechanics; some operations that are integral to certain minimisation procedures (such as matrix inversion) are trivial for small systems but formidable for systems containing thousands of atoms. Quantum mechanics and molecular mechanics also require different amounts of computational effort to calculate the energies and the derivatives of the various configurations. Thus an algorithm that takes many steps may be appropriate for molecular mechanics but inappropriate for quantum mechanics.

Most minimisation algorithms can only go downhill on the energy surface and so they can only locate the minimum that is nearest (in a downhill sense) to the starting point. Thus, Figure 5.3 shows a schematic energy surface and the minima that would be obtained starting from three points A, B and C. The minima can be considered to correspond to the locations where a ball rolling on the energy surface under the influence of gravity would come to rest. To locate more than one minimum or to locate the global energy minimum we therefore usually require a means of generating different starting points, each of which is then minimised. Some specialised minimisation methods can make uphill moves to seek out minima lower in energy than the nearest one, but no algorithm has yet proved capable of locating the

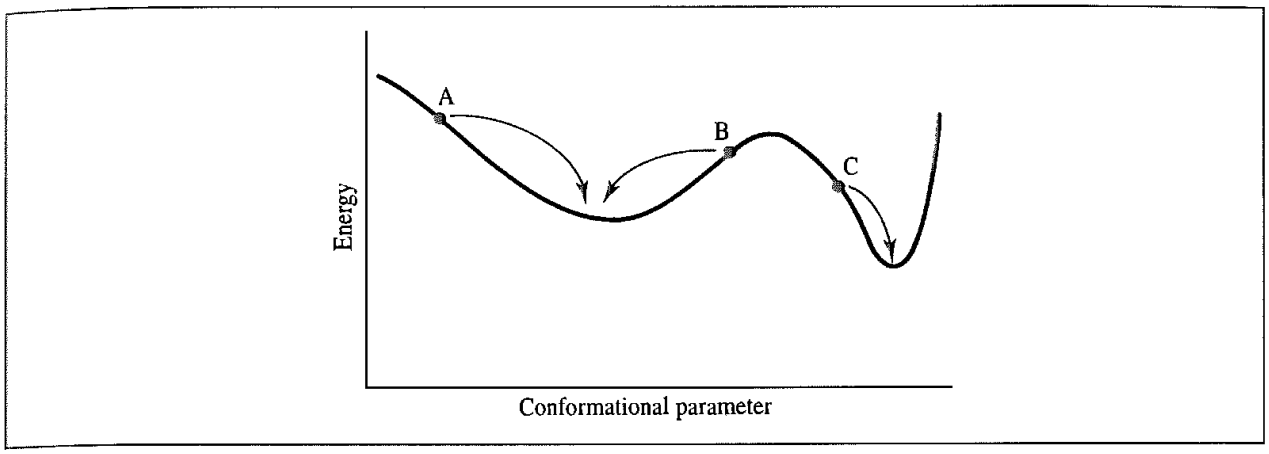

Fig 5 3: A schematic one-dimensional energy surface Minimisation methods move downhill to the nearest minimum The statistical weight of the narrow, deep minimum may be less than a broad minimum which is higher in energy

global energy minimum from an arbitrary starting position. The shape of the energy surface may be important if one wishes to calculate the relative populations of the various minimum energy structures. For example, a deep and narrow minimum may be less highly populated than a broad minimum that is higher in energy as the vibrational energy levels will be more widely spaced in the deeper minimum and so less accessible. For this reason the global energy minimum may not be the most highly populated minimum. In any case, the 'active' structure (e.g. the biologically active conformation of a drug molecule) may not correspond to the global minimum, or to the most highly populated conformation, or even to a minimum energy structure at all.

The input to a minimisation program consists of a set of initial coordinates for the system. The initial coordinates may come from a variety of sources. They may be obtained from an experimental technique, such as X-ray crystallography or NMR. In other cases a theoretical method is employed, such as a conformational search algorithm. A combination of experimental and theoretical approaches may also be used. For example, to study the behaviour of a protein in water one may take an X-ray structure of the protein and immerse it in a solvent 'bath', where the coordinates of the solvent molecules have been obtained from a Monte Carlo or molecular dynamics simulation.

5.1.2 Derivatives

In order to use a derivative minimisation method it is obviously necessary to be able to calculate the derivatives of the energy with respect to the variables (i.e. the Cartesian or internal coordinates, as appropriate). Derivatives may be obtained either analytically or numerically. The use of analytical derivatives is preferable as they are exact, and because they can be calculated more quickly; if only numerical derivatives are available then it may be more effective to use a non-derivative minimisation algorithm. The problems of calculating analytical derivatives with quantum mechanics and molecular mechanics were discussed in Sections 3.4.3 and 4.16, respectively.

Nevertheless, under some circumstances it is necessary to use numerical derivatives. These can be calculated as follows. If one of the coordinates x_i is changed by a small change (δx_i) and the energy for the new arrangement is computed then the derivative $\partial E/\partial x_i$ is obtained by dividing the change in energy (δE) by the change in coordinate ($\delta E/\delta x_i$). This strictly gives the derivative at the mid-point between the two points x_i and $x_i + \delta x_i$. A more accurate value of the derivative at the point x_i may be obtained (at the cost of an additional energy calculation) by evaluating the energy at two points, $x_i + \delta x_i$ and $x_i - \delta x_i$. The derivative is then obtained by dividing the difference in the energies by $2\delta x_i$.

5.2 Non-derivative Minimisation Methods

5.2.1 The Simplex Method

A *simplex* is a geometrical figure with $M + 1$ interconnected vertices, where M is the dimensionality of the energy function. For a function of two variables the simplex is thus triangular in shape. A tetrahedral simplex is used for a function of three variables and so for an energy function of $3N$ Cartesian coordinates the simplex will have $3N + 1$ vertices; if internal coordinates are used then the simplex will have $3N - 5$ vertices. Each vertex corresponds to a specific set of coordinates for which an energy can be calculated. For our function $f(x, y) = x^2 + 2y^2$ the simplex method would use a triangular simplex.

The simplex algorithm locates a minimum by moving around on the potential energy surface in a fashion that has been likened to the motion of an amoeba. Three basic kinds of move are possible. The most common type of move is a reflection of the vertex with the highest value through the opposite face of the simplex, in an attempt to generate a new point that has a lower value. If this new point is lower in energy than any of the other points in the simplex then a 'reflection and expansion' move may be applied. When a 'valley floor' is reached then a reflection move will fail to produce a better point. Under such circumstances the simplex contracts along one dimension from the highest point If this fails to reduce the energy then a third type of move is possible, in which the simplex contracts in all directions, pulling around the lowest point. These three moves are illustrated in Figure 5.4.

To implement the simplex algorithm it is first necessary to generate the vertices of the initial simplex. The initial configuration of the system corresponds to just one of these vertices. The remaining points can be obtained in a variety of ways, but one simple method is to add a constant increment to each coordinate in turn. The energy of the system is calculated at the new point, giving the function value for the relevant vertex.

The simplex method is most useful where the initial configuration of the system is very high in energy, because it rarely fails to find a better solution. However, it can be rather expensive in terms of computer time due to the large number of energy evaluations which are required (merely to generate the initial simplex requires $3N + 1$ energy evaluations). For this reason the simplex method is often used in combination with a different minimisation algorithm: a few steps of the simplex method are used to refine the initial structure and then a more efficient method can take over.

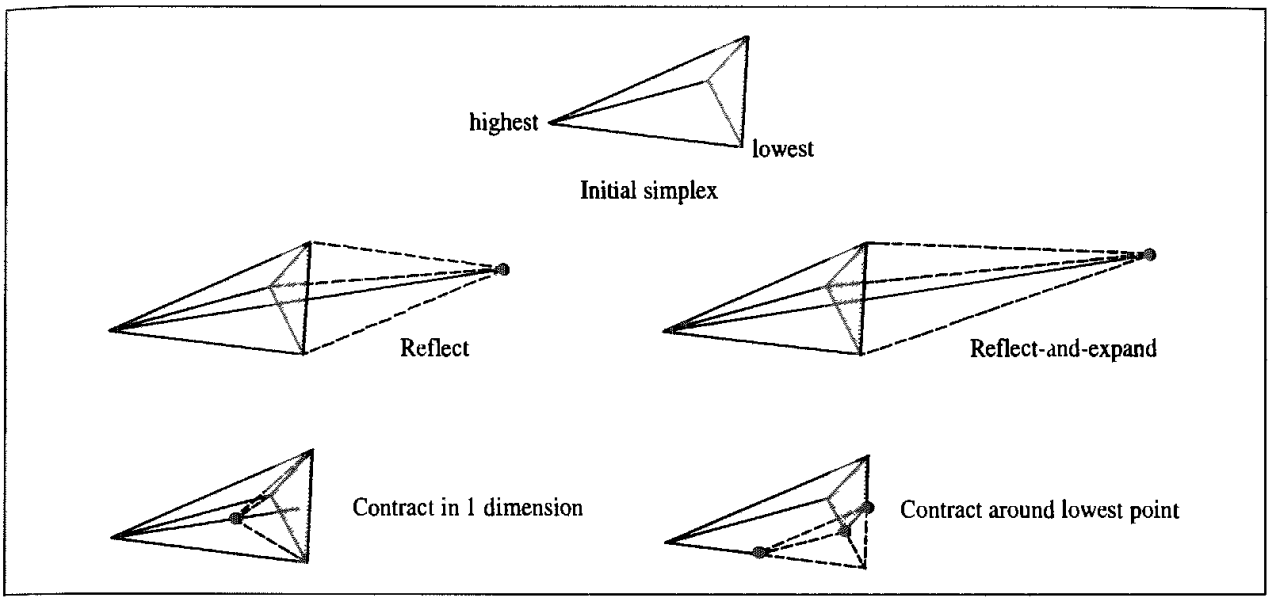

Fig 5.4: The three basic moves permitted to the simplex algorithm (reflection, and its close relation reflect-and-expand; contract in one dimension and contract around the lowest point) (Figure adapted from Press W H, B P Flannery, S A Teukolsky and W T Vetterling 1992. Numerical Recipes in Fortran. *Cambridge, Cambridge University Press.)*

Let us consider the application of the simplex method to our quadratic function, $f = x^2 + 2y^2$ (Figure 5.5). Suppose our initial simplex contains vertices located at the points (9, 9), (11, 9) and (9, 11), which have been generated by adding a constant factor 2 to each of the variables in turn. The values of the function at these points are 243, 283 and 323, respectively. The vertex with the highest function value is at (9, 11) and so in the first iteration this point is reflected through the opposite face of the triangle to generate a point with coordinates (11, 7) and a function value of 219 (we do not use the reflect-and-expand move in our

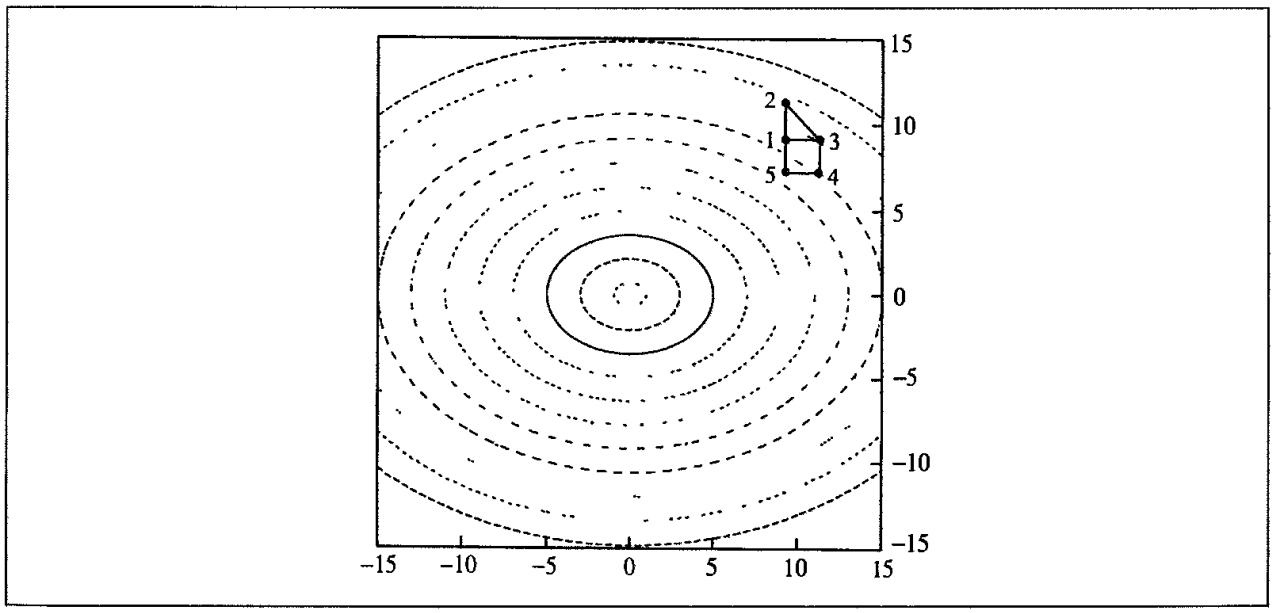

Fig 5.5: The first few steps of the simplex algorithm with the function $x^2 + 2y^2$. The initial simplex corresponds to the triangle 123. Point 2 has the largest value of the function and the next simplex is the triangle 134. The simplex for the third step is 145.

illustration). The highest vertex is now at (11, 9), which is reflected through the opposite face of the simplex to give the point (9, 7), where the function has a value of 179. In fact, for this admittedly artificial problem the simplex algorithm takes more than 30 steps to find a point where the function has a value less than 0.1.

Why does the simplex contain one more vertex than the number of degrees of freedom? The reason is that with fewer than $M + 1$ vertices the algorithm cannot explore the whole energy surface. Suppose we use only a two-vertex simplex to explore our quadratic energy surface. A simplex with just two vertices is a straight line. The only moves that would be possible in this case would be to other points that lie on this line; none of the energy surface away from the line would be explored. Similarly, if we have a function of three variables and restrict the simplex to a triangle then we will only be able to explore the region of space that lies in the same plane as the triangle, whereas the minimum may not lie in this plane.

5.2.2 The Sequential Univariate Method

The simplex method is rarely considered suitable for quantum mechanical calculations, due to the number of energy evaluations that must be performed. The sequential univariate method is a non-derivative method that is considered more appropriate in this case. This method systematically cycles through the coordinates in turn. For each coordinate, two new structures are generated by changing the current coordinate (i.e. $x_i + \delta x_i$ and $x_i + 2\delta x_i$). The energies of these two structures are calculated. A parabola is then fitted through the three points corresponding to the two distorted structures and the original structure. The minimum point in this quadratic function is determined and the coordinate is then changed to the position of the minimum The procedure is illustrated in Figure 5.6. When the changes in all the coordinates are

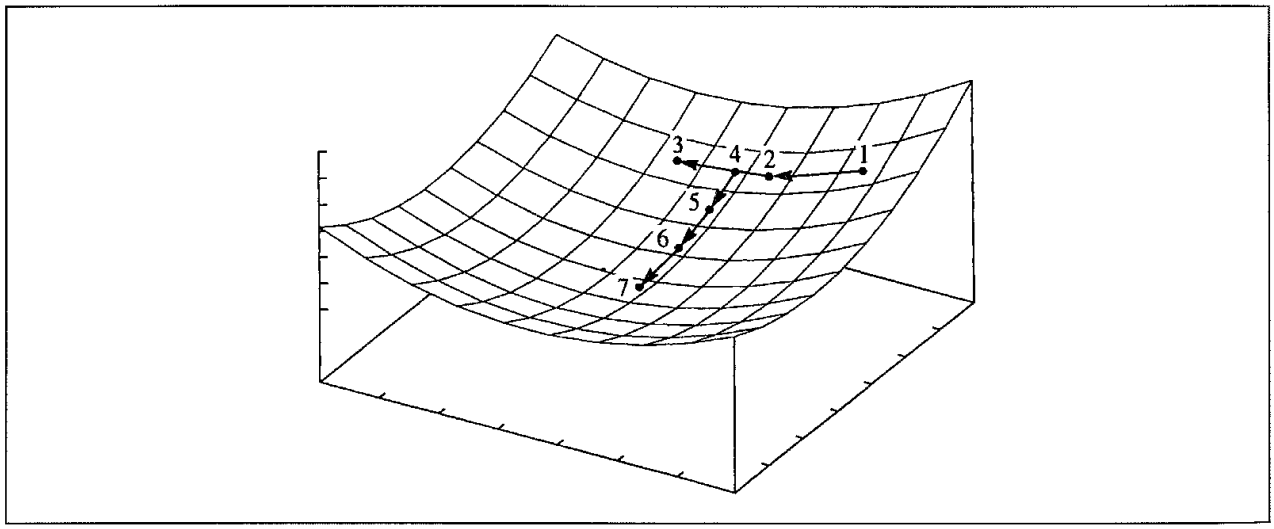

Fig. 5.6. The sequential univariate method Starting at the point labelled 1 two steps are made along one of the coordinates to give points 2 and 3 A parabola is fitted to these three points and the minimum located (point 4) The same procedure is then repeated along the next coordinate (points 5, 6 and 7). (Figure adapted from Schlegel H B 1987. Optimization of Equilibrium Geometries and Transition Structures In Lawley K P (Editor) Ab Initio Methods in Quantum Chemistry - I *New York, John Wiley, pp. 249–286)*

sufficiently small then the minimum is deemed to have been reached, otherwise a new iteration is performed. The sequential invariate method usually requires fewer function evaluations than the simplex method but it can be slow to converge especially if there is strong coupling between two or more of the coordinates or when the energy surface is analogous to a long narrow valley.

5.3 Introduction to Derivative Minimisation Methods

Derivatives provide information that can be very useful in energy minimisation, and derivatives are used by most popular minimisation methods. The direction of the first derivative of the energy (the gradient) indicates where the minimum lies, and the magnitude of the gradient indicates the steepness of the local slope. The energy of the system can be lowered by moving each atom in response to the force acting on it; the force is equal to minus the gradient. Second derivatives indicate the curvature of the function, information that can be used to predict where the function will change direction (i.e. pass through a minimum or some other stationary point).

When discussing derivative methods it is useful to write the function as a Taylor series expansion about the point x_k:

$$\mathscr{V}(x) = \mathscr{V}(x_k) + (x - x_k)\mathscr{V}'(x_k) + (x - x_k)^2\mathscr{V}''(x_k)/2 + \cdots \tag{5.2}$$

For a multidimensional function, the variable x is replaced by the vector $\mathbf{x}$ and matrices are used for the various derivatives. Thus if the potential energy $\mathscr{V}(\mathbf{x})$ is a function of $3N$ Cartesian coordinates, the vector $\mathbf{x}$ will have $3N$ components and $\mathbf{x}_k$ corresponds to the current configuration of the system. $\mathscr{V}'(\mathbf{x}_k)$ is a $3N \times 1$ matrix (i.e. a vector), each element of which is the partial derivative of $\mathscr{V}$ with respect to the appropriate coordinate, $\partial\mathscr{V}/\partial x_i$. We will also write the gradient at the point k as $\mathbf{g}_k$. Each element (i,j) of the matrix $\mathscr{V}''(\mathbf{x}_k)$ is the partial second derivative of the energy function with respect to the two coordinates x_i and x_j, $\partial^2\mathscr{V}/\partial x_i\partial x_j$. $\mathscr{V}''(\mathbf{x}_k)$ is thus of dimension $3N \times 3N$ and is known as the *Hessian* matrix or the *force constant* matrix. The Taylor series expansion can be written in the following form for the multidimensional case:

$$\mathscr{V}(\mathbf{x}) = \mathscr{V}(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)\mathscr{V}'(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^{\mathrm{T}} \cdot \mathscr{V}''(\mathbf{x}_k) \cdot (\mathbf{x} - \mathbf{x}_k)/2 + \cdots \tag{5.3}$$

The energy functions used in molecular modelling are rarely quadratic and so the Taylor series expansion, Equation (5.3), can only be considered an approximation. There are two important consequences of this. The first consequence is that the performance of a given minimisation method will not be as good for a molecular mechanics or quantum mechanics energy surface as it is for a pure quadratic function. As we shall see, a second derivative method such as the Newton–Raphson algorithm can locate the minimum in a single step for a purely quadratic function, but several iterations are usually required for a typical molecular modelling energy function. The second consequence is that, far from the minimum, the harmonic approximation is a poor one and some of the less robust methods will fail, even though they may work very well close to a minimum, where the harmonic approximation is more valid. For this reason it is important to choose the minimisation

protocol with care, possibly using a robust (but perhaps inefficient) method at first, and then a less robust but more efficient method.

The derivative methods can be classified according to the highest-order derivative used. First-order methods use the first derivatives (i.e. the gradients) whereas second-order methods use both first and second derivatives. The simplex method can thus be considered a zeroth-order method as it does not use any derivatives.

5.4 First-order Minimisation Methods

Two first-order minimisation algorithms that are frequently used in molecular modelling are the method of *steepest descents* and the *conjugate gradient* method. These gradually change the coordinates of the atoms as they move the system closer and closer to the minimum point. The starting point for each iteration (k) is the molecular configuration obtained from the previous step, which is represented by the multidimensional vector $\mathbf{x}_{k-1}$. For the first iteration the starting point is the initial configuration of the system provided by the user, the vector $\mathbf{x}_1$.

5.4.1 The Steepest Descents Method

The steepest descents method moves in the direction parallel to the net force, which in our geographical analogy corresponds to walking straight downhill. For $3N$ Cartesian coordinates this direction is most conveniently represented by a $3N$-dimensional unit vector, $\mathbf{s}_k$. Thus:

$$\mathbf{s}_k = -\mathbf{g}_k/|\mathbf{g}_k| \tag{5.4}$$

Having defined the direction along which to move it is then necessary to decide how far to move along the gradient. Consider the two-dimensional energy surface of Figure 5.7. The gradient direction from the starting point is along the line indicated. If we imagine a cross-section through the surface along the line, the function will pass through a minimum and then increase, as shown in the figure. We can choose to locate the minimum point by performing a *line search* or we can take a step of arbitrary size along the direction of the force

5.4.2 Line Search in One Dimension

The purpose of a line search is to locate the minimum along a specified direction (i.e. along a line through the multidimensional space). The first stage of the line search is to *bracket* the minimum. This entails finding three points along the line such that the energy of the middle point is lower than the energy of the two outer points. If three such points can be found, then at least one minimum must lie between the two outer points. An iterative procedure can then be used to decrease the distance between the three points, gradually restricting the minimum to an even smaller region. This is conceptually an easy process but it may require a considerable number of function evaluations, making it computationally expensive. An alternative is to fit a function such as a quadratic to the three points. Differentiation of the

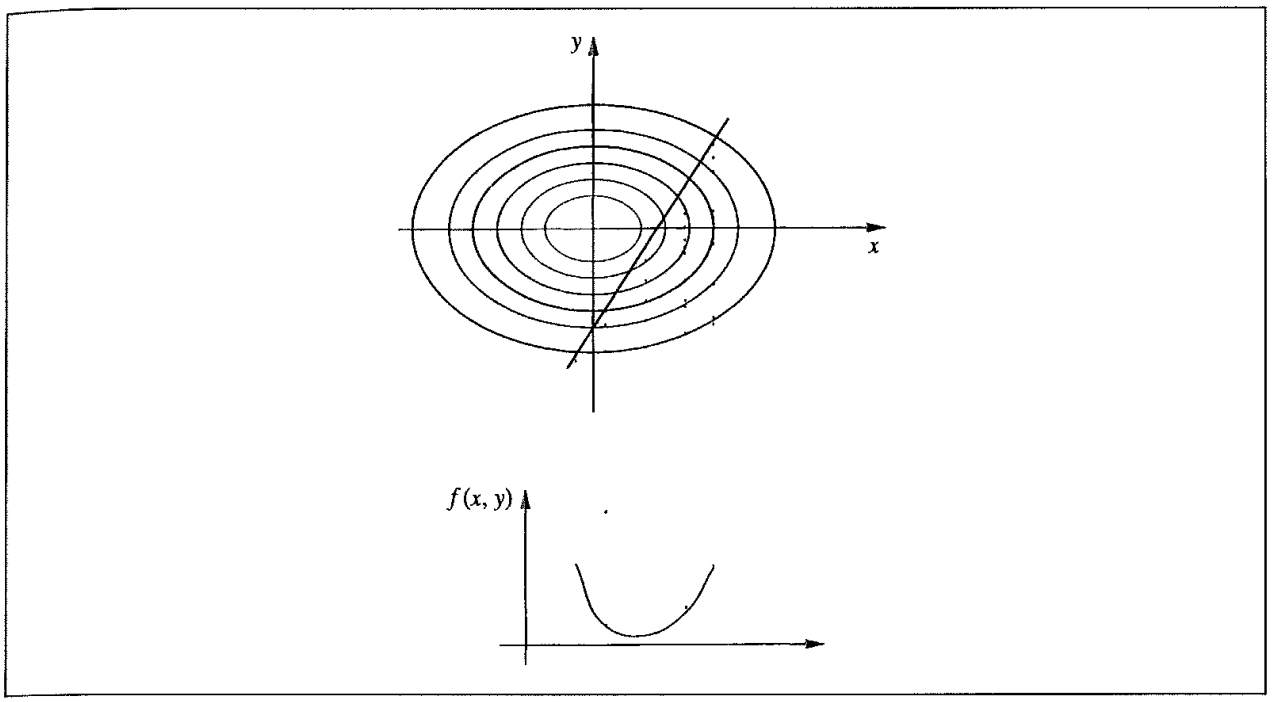

Fig 5 7 A line search is used to locate the minimum in the function in the direction of the gradient.

fitted function enables an approximation to the minimum along the line to be identified analytically. A new function can then be fitted to give a better estimate, as shown in Figure 5.8. Higher-order polynomials may give a better fit to the bracketing points but these can give incorrect interpolations when used with functions that change sharply in the bracketed region.

The gradient at the minimum point obtained from the line search will be perpendicular to the previous direction. Thus, when the line search method is used to locate the minimum along the gradient then the next direction in the steepest descents algorithm will be orthogonal to the previous direction (i.e. $\mathbf{g}_k \cdot \mathbf{g}_{k-1} = 0$).

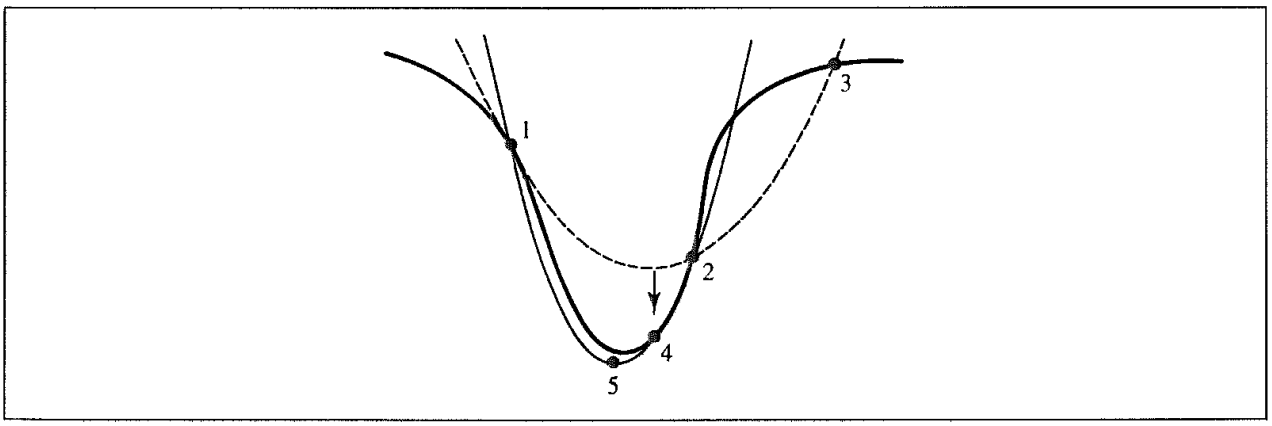

Fig 5 8: The minimum in a line search may be found more effectively by fitting an analytical function such as a quadratic to the initial set of three points (1, 2 and 3) A better estimate of the minimum can then be found by fitting a new function to the points 1, 2 and 4 and finding its minimum. (Figure adapted from Press W H, B P Flannery, S A Teukolsky and W T Vetterling 1992 Numerical Recipes in Fortran *Cambridge, Cambridge University Press)*

5.4.3 Arbitrary Step Approach

As the line search may itself be computationally demanding we could obtain the new coordinates by taking a step of arbitrary length along the gradient unit vector $\mathbf{s}_k$. The new set of coordinates after step k would then be given by the equation:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{s}_k \tag{5.5}$$

λ_k is the *step size*. In most applications of the steepest descents algorithm in molecular modelling the step size initially has a predetermined default value. If the first iteration leads to a reduction in energy, the step size is increased by a multiplicative factor (e.g. 1.2) for the second iteration. This process is repeated so long as each iteration reduces the energy. When a step produces an increase in energy, it is assumed that the algorithm has leapt across the valley which contains the minimum and up the slope on the opposite face The step size is then reduced by a multiplicative factor (e.g. 0.5). The step size depends upon the nature of the energy surface; for a flat surface large step sizes would be appropriate but for a narrow, twisting gully a much smaller step would be more suitable. The arbitrary step method may require more steps to reach the minimum but it can often require fewer function evaluations (and thus less computer time) than the more rigorous line search approach.

The steepest descents method works as follows for our trial function, $f(x,y) = x^2 + 2y^2$. Differentiating the function gives $df = 2x\,dx + 4y\,dy$ and so the gradient at any point $(x.y)$ equals $4y/2x$. The direction of the first move from the point (9 0,9 0) is (−18 0, −36.0) and the equation of the line along which the search is performed is $y = 2x - 9$. The minimum of the function along thus line can be obtained using Lagrange multipliers (see Section 1.10.5) and is at $(4.0, -1.0)$. The direction of the next move is the vector $(-8,4)$ and the next line search is performed along the line $y = -0.5x + 1$. The minimum point along thus line is $(2/3, 2/3)$ where the function has the value $4/3$. The third point found by the steepest descents method is at $(0.296, -0.074)$ where the function has the value 0.099. These moves are illustrated in Figure 5.9.

The direction of the gradient is determined by the largest interatomic forces and so steepest descents is a good method for relieving the highest-energy features in an initial configuration. The method is generally robust even when the starting point is far from a minimum, where the harmonic approximation to the energy surface is often a poor assumption. However, it suffers from the problem that many small steps will be performed when proceeding down a long narrow valley. The steepest descents method is forced to make a right-angled turn at each point, even though that might not be the best route to the minimum. The path oscillates and continually overcorrects itself, as illustrated in Figure 5.10; later steps reintroduce errors that were corrected by earlier moves.

5.4.4 Conjugate Gradients Minimisation

The conjugate gradients method produces a set of directions which does not show the oscillatory behaviour of the steepest descents method in narrow valleys. In the steepest descents method both the gradients and the direction of successive steps are orthogonal.

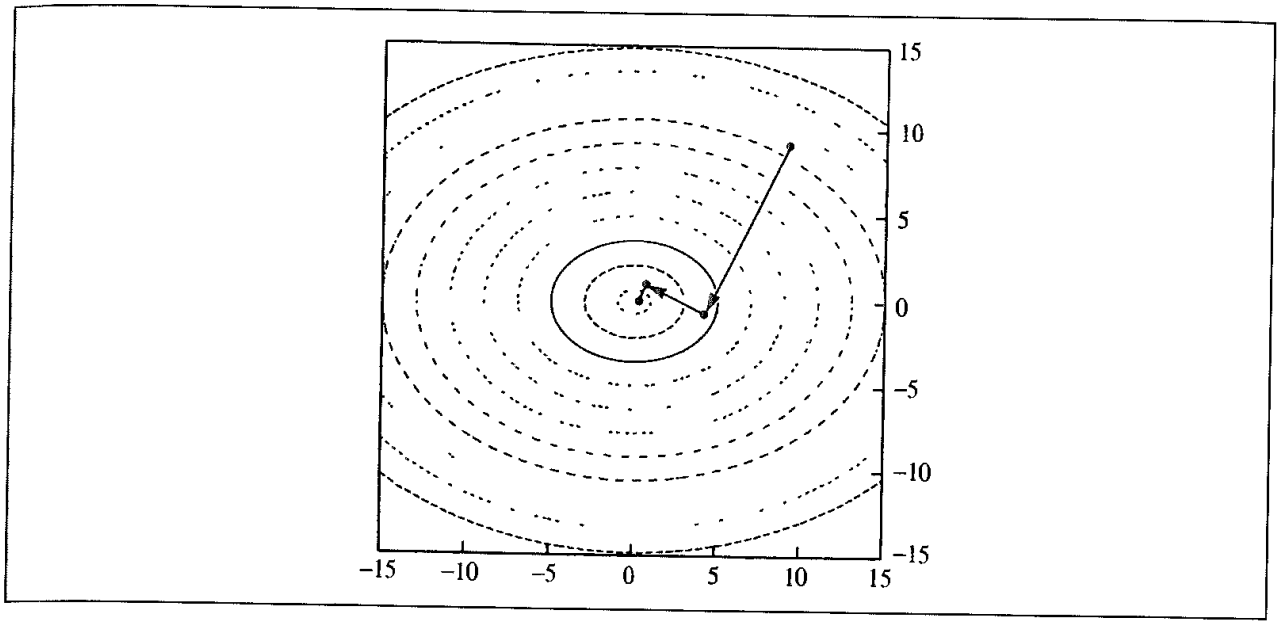

Fig 5 9. Application of steepest descents to the function $x^2 + 2y^2$.

In conjugate gradients, the gradients at each point are orthogonal but the directions are *conjugate* (indeed, the method is more properly called the conjugate directions method). A set of conjugate directions has the property that for a quadratic function of M variables, the minimum will be reached in M steps The conjugate gradients method moves in a direction $\mathbf{v}_k$ from point $\mathbf{x}_k$ where $\mathbf{v}_k$ is computed from the gradient at the point and the previous direction vector $\mathbf{v}_{k-1}$:

$$\mathbf{v}_k = -\mathbf{g}_k + \gamma_k \mathbf{v}_{k-1} \tag{5.6}$$

γ_k is a scalar constant given by

$$\gamma_k = \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}} \tag{5 7}$$

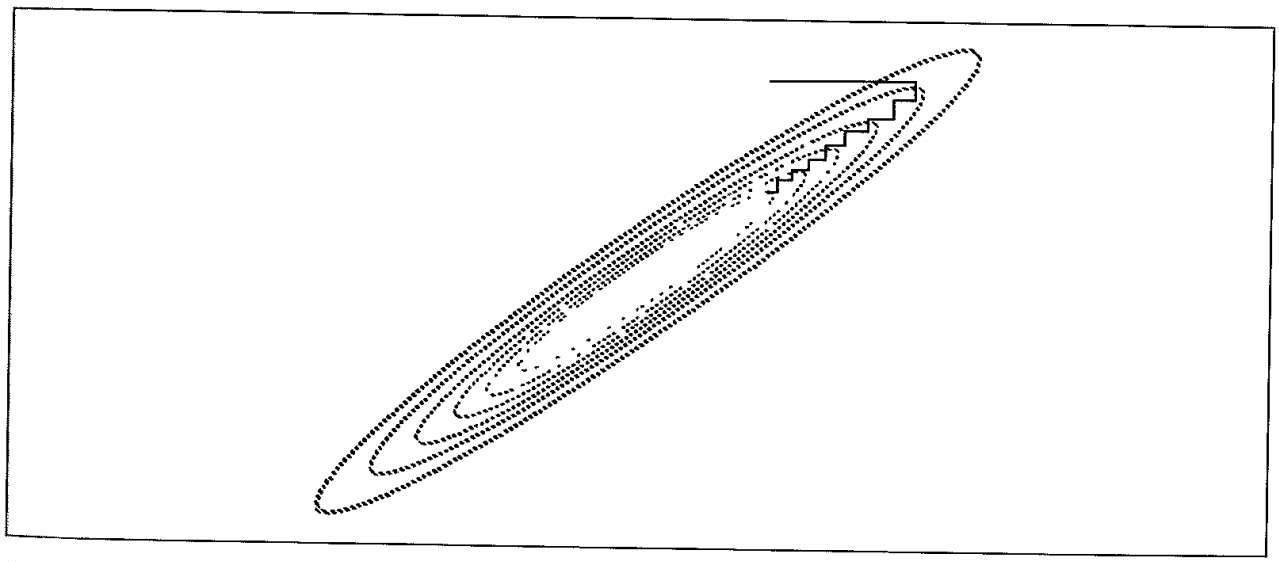

Fig 5 10 The steepest descents method can give undesirable behaviour in a long narrow valley

In the conjugate gradients method all of the directions and gradients satisfy the following relationships:

$$\mathbf{g}_i \cdot \mathbf{g}_j = 0 \tag{5.8}$$

$$\mathbf{v}_i \cdot \mathbf{H}_{ij} \cdot \mathbf{v}_j = 0 \tag{5.9}$$

$$\mathbf{g}_i \cdot \mathbf{v}_j = 0 \tag{5.10}$$

Clearly Equation (5.6) can only be used from the second step onwards and so the first step in the conjugate gradients method is the same as the steepest descents (i.e. in the direction of the gradient). The line search method should ideally be used to locate the one-dimensional minimum in each direction to ensure that each gradient is orthogonal to all previous gradients and that each direction is conjugate to all previous directions. However, an arbitrary step method is also possible.

The conjugate gradients method deals with our simple quadratic function $f(x, y) = x^2 + 2y^2$ as follows. From the initial point (9, 9) we move to the same point as in steepest descents, (4,−1). To find the direction of the next move, we first determine the negative gradient at the current point. This is the vector (−8, 4). This is then combined with the vector corresponding to minus the gradient at the initial point, (−18,−36) multiplied by γ:

$$\mathbf{v}_k = \begin{pmatrix} -8 \\ 4 \end{pmatrix} + \frac{(-8)^2 + (4)^2}{(-18)^2 + (-36)^2} \begin{pmatrix} -18 \\ -36 \end{pmatrix} = \begin{pmatrix} -80/9 \\ +20/9 \end{pmatrix} \tag{5.11}$$

To locate the second point we therefore need to perform a line search along the line with gradient −1/4 that passes through the point (4,−1). The minimum along this line is at the origin, at the true minimum of the function. The conjugate gradients method thus locates the exact minimum of the function exactly in just two moves, as illustrated in Figure 5.11.

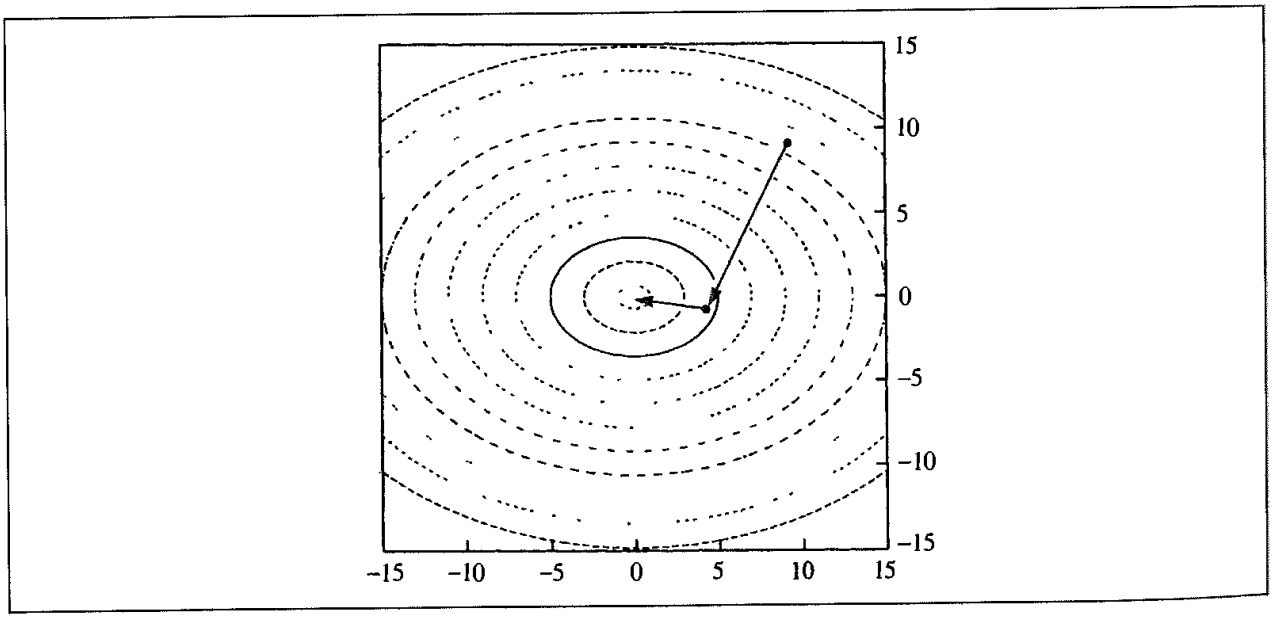

Fig 5.11. Application of conjugate gradients method to the function $x^2 + 2y^2$.

Several variants of the conjugate gradients method have been proposed. The formulation given in Equation (5.7) is the original Fletcher–Reeves algorithm. Polak and Ribiere proposed an alternative form for the scalar constant γ_k:

$$\gamma_k = \frac{(\mathbf{g}_k - \mathbf{g}_{k-1}) \cdot \mathbf{g}_k}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}} \tag{5.12}$$

For a purely quadratic function the Polak–Ribiere method is identical to the Fletcher–Reeves algorithm as all gradients will be orthogonal. However, most functions of interest, including those used in molecular modelling, are at best only approximately quadratic. Polak and Riviere claimed that their method performed better than the original Fletcher–Reeves algorithm, at least for the functions that they examined.

5.5 Second Derivative Methods: The Newton–Raphson Method

Second-order methods use not only the first derivatives (i.e. the gradients) but also the second derivatives to locate a minimum. Second derivatives provide information about the curvature of the function. The *Newton–Raphson* method is the simplest second-order method. Recall our Taylor series expansion about the point x_k, Equation (5.2):

$$\mathscr{V}(x) = \mathscr{V}(x_k) + (x - x_k)\mathscr{V}'(x_k) + (x - x_k)^2\mathscr{V}''(x_k)/2 + \cdots \tag{5.13}$$

The first derivative of $\mathscr{V}(x)$ is:

$$\mathscr{V}'(x) = x\mathscr{V}'(x_k) + (x - x_k)\mathscr{V}''(x_k) \tag{5.14}$$

If the function is purely quadratic, the second derivative is the same everywhere, and so $\mathscr{V}''(x) = \mathscr{V}''(x_k)$.

At the minimum $(x = x^*)\mathscr{V}'(x^*) = 0$ and so

$$x^* = x_k - \mathscr{V}'(x_k)/\mathscr{V}''(x_k) \tag{5.15}$$

For a multidimensional function: $\mathbf{x}^* = \mathbf{x}_k - \mathscr{V}'(\mathbf{x}_k)\mathscr{V}''^{-1}(\mathbf{x}_k)$.

$\mathscr{V}''^{-1}(\mathbf{x}_k)$ is the inverse Hessian matrix of second derivatives, which, in the Newton–Raphson method, must therefore be inverted. This can be computationally demanding for systems with many atoms and can also require a significant amount of storage. The Newton–Raphson method is thus more suited to small molecules (usually less than 100 atoms or so). For a purely quadratic function the Newton–Raphson method finds the minimum in one step from any point on the surface, as we will now show for our function $f(x, y) = x^2 + 2y^2$.

The Hessian matrix for this function is:

$$\mathbf{f}'' = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \tag{5.16}$$

The inverse of this matrix is:

$$\mathbf{f}''^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/4 \end{pmatrix} \tag{5.17}$$

The minimum is obtained using Equation (5.15):

$$\mathbf{x}^* = \begin{pmatrix} 9 \\ 9 \end{pmatrix} - \begin{pmatrix} 1/2 & 0 \\ 0 & 1/4 \end{pmatrix} \begin{pmatrix} 18 \\ 36 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5.18}$$

In practice, of course, the surface is only quadratic to a first approximation and so a number of steps will be required, at each of which the Hessian matrix must be calculated and inverted. The Hessian matrix of second derivatives must be *positive definite* in a Newton–Raphson minimisation. A positive definite matrix is one for which all the eigenvalues are positive. When the Hessian matrix is not positive definite then the Newton–Raphson method moves to points (e.g. saddle points) where the energy increases. In addition, far from a minimum the harmonic approximation is not appropriate and the minimisation can become unstable. One solution to this problem is to use a more robust method to get near to the minimum (i.e. where the Hessian is positive definite) before applying the Newton–Raphson method.

5.5.1 Variants on the Newton–Raphson Method

There are a number of variations on the Newton–Raphson method, many of which aim to eliminate the need to calculate the full matrix of second derivatives. In addition, a family of methods called the quasi-Newton methods require only first derivatives and gradually construct the inverse Hessian matrix as the calculation proceeds. One simple way in which it may be possible to speed up the Newton–Raphson method is to use the same Hessian matrix for several successive steps of the Newton–Raphson algorithm with only the gradients being recalculated at each iteration.

A widely used algorithm is the *block-diagonal Newton–Raphson* method in which just one atom is moved at each iteration. Consequently all terms of the form $\partial^2 \mathcal{V} / \partial x_i \partial x_j$, where i and j refer to the Cartesian coordinates of atoms other than the atom being moved, will be zero. This only leaves those terms which involve the coordinates of the atom being moved and so reduces the problem to the trivial one of inverting a 3×3 matrix. However, the block-diagonal approach can be less efficient when the motions of some atoms are closely coupled, such as the concerted movements of connected atoms in a phenyl ring.

5.6 Quasi-Newton Methods

Calculation of the inverse Hessian matrix can be a potentially time-consuming operation that represents a significant drawback to the 'pure' second derivative methods such as Newton–Raphson. Moreover, one may not be able to calculate analytical second derivatives, which are preferable. The quasi-Newton methods (also known as variable metric methods) gradually build up the inverse Hessian matrix in successive iterations. That is, a sequence of

matrices $\mathbf{H}_k$ is constructed that has the property

$$\lim_{k\to\infty} \mathbf{H}_k = \mathscr{V}''^{-1} \tag{5 19}$$

At each iteration k, the new positions $\mathbf{x}_{k+1}$ are obtained from the current positions $\mathbf{x}_k$, the gradient $\mathbf{g}_k$ and the current approximation to the inverse Hessian matrix $\mathbf{H}_k$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}_k\mathbf{g}_k \tag{5 20}$$

This formula is exact for a quadratic function, but for 'real' problems a line search may be desirable. This line search is performed along the vector $\mathbf{x}_{k+1} - \mathbf{x}_k$. It may not be necessary to locate the minimum in the direction of the line search very accurately, at the expense of a few more steps of the quasi-Newton algorithm. For quantum mechanics calculations the additional energy evaluations required by the line search may prove more expensive than using the more approximate approach. An effective compromise is to fit a function to the energy and gradient at the current point $\mathbf{x}_k$ and at the point $\mathbf{x}_{k+1}$ and determine the minimum in the fitted function.

Having moved to the new positions $\mathbf{x}_{k+1}$, $\mathbf{H}$ is updated from its value at the previous step according to a formula depending upon the specific method being used. The methods of Davidon–Fletcher–Powell (DFP), Broyden–Fletcher–Goldfarb–Shanno (BFGS) and Murtaugh–Sargent (MS) are commonly encountered, but there are many others. These methods converge to the minimum, for a quadratic function of M variables, in M steps. The DFP formula is:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{x}_{k+1} - \mathbf{x}_k) \otimes (\mathbf{x}_{k-1} - \mathbf{x}_k)}{(\mathbf{x}_{k+1} - \mathbf{x}_k)\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)} - \frac{[\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)] \otimes [\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)]}{(\mathbf{g}_{k+1} - \mathbf{g}_k)\cdot\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)} \tag{5.21}$$

The symbol $\otimes$ when interposed between two vectors means that a matrix is to be formed. The ijth element of the matrix $\mathbf{u}\otimes\mathbf{v}$ is obtained by multiplying $\mathbf{u}_i$ by $\mathbf{v}_j$.

The BFGS formula differs from the DFP equation by an additional term:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{x}_{k+1} - \mathbf{x}_k) \otimes (\mathbf{x}_{k+1} - \mathbf{x}_k)}{(\mathbf{x}_{k+1} - \mathbf{x}_k)\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)} - \frac{[\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)] \otimes [\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)]}{(\mathbf{g}_{k+1} - \mathbf{g}_k)\cdot\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)} + [(\mathbf{g}_{k+1} - \mathbf{g}_k)\cdot\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)]\mathbf{u}\otimes\mathbf{u} \tag{5 22}$$

where

$$\mathbf{u} = \frac{(\mathbf{x}_{k+1} - \mathbf{x}_k)}{(\mathbf{x}_{k+1} - \mathbf{x}_k)\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)} - \frac{[\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)]}{(\mathbf{g}_{k+1} - \mathbf{g}_k)\cdot\mathbf{H}_k\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)} \tag{5.23}$$

The MS formula is:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{[(\mathbf{x}_{k+1} - \mathbf{x}_k) - \mathbf{H}_k(\mathbf{g}_{k+1} - \mathbf{g}_k)] \otimes [(\mathbf{x}_{k+1} - \mathbf{x}_k) - \mathbf{H}_k(\mathbf{g}_{k+1} - \mathbf{g}_k)]}{[(\mathbf{x}_{k+1} - \mathbf{x}_k) - \mathbf{H}_k(\mathbf{g}_{k+1} - \mathbf{g}_k)]\cdot(\mathbf{g}_{k+1} - \mathbf{g}_k)} \tag{5 24}$$

All of these methods use just the new and current points to update the inverse Hessian. The default algorithm used in the Gaussian series of molecular orbital programs [Schlegel 1982] makes use of more of the previous points to construct the Hessian (and thence the inverse Hessian), giving better convergence properties. Another feature of this method is its use

of a quartic polynomial that is guaranteed to have just one local minimum in the line search. The DFP, BFGS and MS methods can also be used with numerical derivatives, but alternative approaches may be more effective under such circumstances.

The matrix **H** is often initialised to the unit matrix **I**. The performance of the quasi-Newton algorithms can be improved by using a better estimate of the inverse Hessian than just the unit matrix. The unit matrix gives no information about the bonding in the system, nor does it identify any coupling between the various degrees of freedom. For example, a molecular mechanics calculation can be used to provide an initial guess to **H** prior to a quantum mechanical calculation. Alternatively the matrix can be obtained from a quantum mechanical calculation at a lower level of theory (e.g. semi-empirical or with a smaller basis set).

5.7 Which Minimisation Method Should I Use?

The choice of minimisation algorithm is dictated by a number of factors, including the storage and computational requirements, the relative speeds with which the various parts of the calculation can be performed, the availability of analytical derivatives and the robustness of the method. Thus, any method that requires the Hessian matrix to be stored (let alone its inverse calculated) may present memory problems when applied to systems containing thousands of atoms. Calculations on systems of this size are invariably performed using molecular mechanics, and so the steepest descents and the conjugate gradients methods are very popular here. For molecular mechanics calculations on small molecules, the Newton–Raphson method may be used, although this algorithm can have problems with structures that are far from a minimum. For this reason it is usual to perform a few steps of minimisation using a more robust method such as the simplex or steepest descents before applying the Newton–Raphson algorithm Analytical expressions for both first and second derivatives are available for most of the terms found in common force fields.

The performance of the steepest descents and conjugate gradients methods is contrasted in the following example. A model of the antibiotic netropsin (Figure 5.12) bound to DNA was constructed using an automated docking program. This initial model was then subjected to two stages of minimisation. In the first stage, the aim was to produce a structure that did not

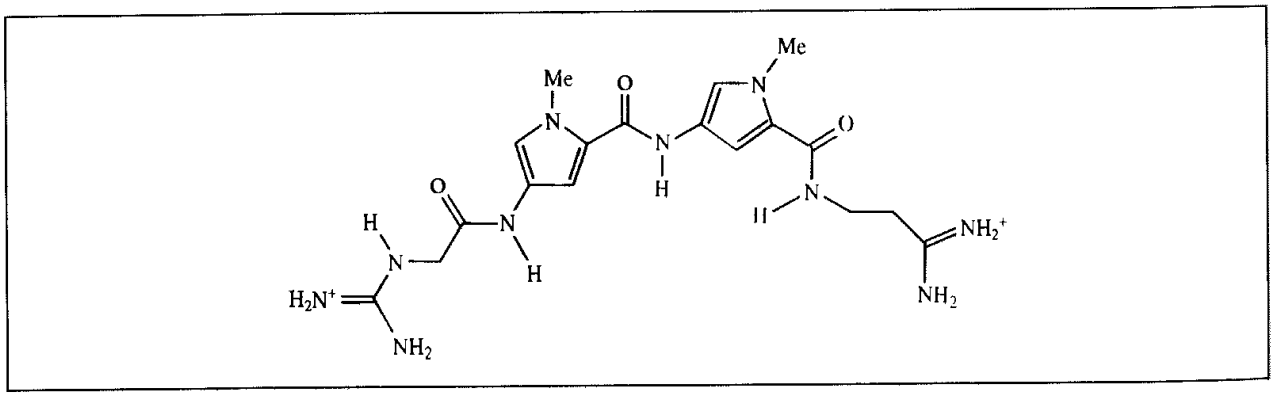

Fig 5.12· The DNA inhibitor netropsin

	Initial refinement (Av. gradient <1 kcal Å^{-2})		Stringent minimisation (Av. gradient <0.1 kcal Å^{-2})	
Method	**CPU time (s)**	**Number of iterations**	**CPU time (s)**	**Number of iterations**
Steepest descents	67	98	1405	1893
Conjugate gradients	149	213	257	367

Table 5.1 A comparison of the steepest descents and conjugate gradients methods for an initial refinement and a stringent minimisation.

have any significant high-energy interactions. The structure was then further minimised to give a structure much closer to the minimum. The results are shown in Table 5.1.

This study shows that the steepest descent method can actually be superior to conjugate gradients when the starting structure is some way from the minimum. However, conjugate gradients is much better once the initial strain has been removed.

Quantum mechanical calculations are restricted to systems with relatively small numbers of atoms, and so storing the Hessian matrix is not a problem. As the energy calculation is often the most time-consuming part of the calculation, it is desirable that the minimisation method chosen takes as few steps as possible to reach the minimum. For many levels of quantum mechanics theory analytical first derivatives are available. However, analytical second derivatives are only available for a few levels of theory and can be expensive to compute. The quasi-Newton methods are thus particularly popular for quantum mechanical calculations.

When using internal coordinates in a quantum mechanical minimisation it can be important to use an appropriate Z-matrix as input. For many systems the Z-matrix can often be written in many different ways as there are many combinations of internal coordinates. There should be no strong coupling between the coordinates. *Dummy atoms* can often help in the construction of an appropriate Z-matrix. A dummy atom is used solely to define the geometry and has no nuclear charge and no basis functions. A simple example of the use of dummy atoms is for a linear molecule such as HN_3, where the angle of 180° would cause problems. The geometry of this molecule can be defined using a dummy atom as illustrated in Figure 5.13; the associated Z-matrix for this system would be:

```
1   N
2   N   1   RN1N2
3   X   1   1.0      2   90.0
4   N   1   RN1N4    3   AN4N1X   2   180.0
5   H   4   RN4H     1   AHN4N1   3   180.0
```

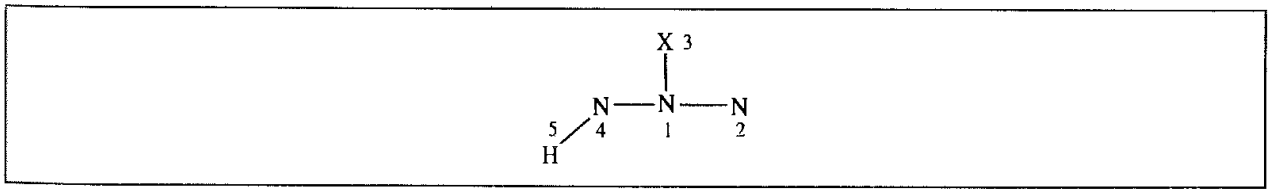

Fig 5 13· Internal coordinates of HN_3 molecule defined using dummy atom X

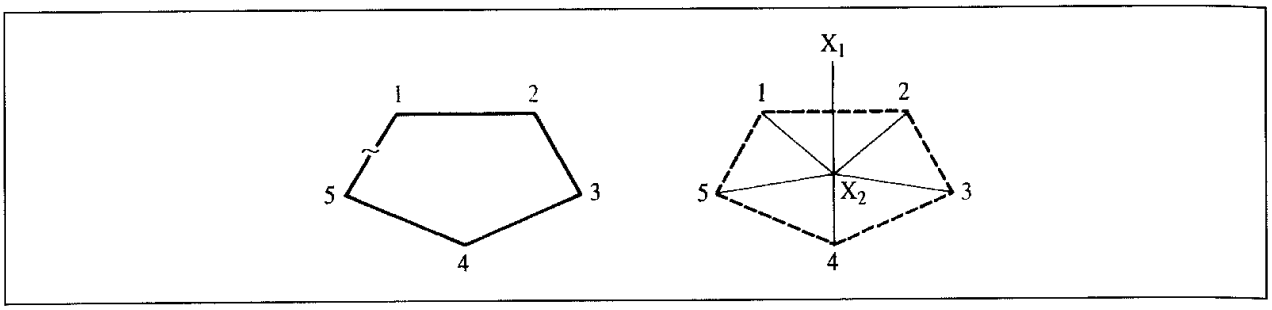

Fig 5 14· The ring closure bond between atoms 1 and 5 would be strongly coupled to the other internal coordinates (left) unless dummy atoms are used to define the Z-matrix (right)

Strong coupling between coordinates can give long 'valleys' in the energy surface, which may also present problems. Care must be taken when defining the Z-matrix for cyclic systems in particular. The natural way to define a cyclic compound would be to number the atoms sequentially around the ring. However, this would then mean that the ring closure bond will be very strongly coupled to all of the other bonds, angles and torsion angles (Figure 5.14). A better definition uses a dummy atom placed at the centre of the ring (Figure 5.14). Some quantum mechanics programs are able to convert the input coordinates (be they Cartesian or internal) into the most efficient set for minimisation so removing from the user the problems of trying to decide what is an appropriate set of internal coordinates. For energy minimisations redundant internal coordinates have been shown to give significant improvements in efficiency compared with Cartesian coordinates or non-redundant internal coordinates, especially for flexible and polycyclic systems [Peng *et al.* 1996]. The redundant internal coordinates employed generally comprise the bond lengths, angles and torsion angles in the system. These methods obviously also require the means to interconvert between the internal coordinate representation and the Cartesian coordinates that are often used as input and desired as output. Of particular importance is the need to transform energy derivatives and the Hessian matrices (if appropriate).

5.7.1 Distinguishing Between Minima, Maxima and Saddle Points

A configuration at which all the first derivatives are zero need not necessarily be a minimum point; this condition holds at both maxima and saddle points as well. From simple calculus we know that the second derivative of a function of one variable, $f'(x)$ is positive at a minimum and negative at a maximum. It is necessary to calculate the eigenvalues of the Hessian matrix to distinguish between minima, maxima and saddle points. At a minimum point there will be six zero and $3N-6$ positive eigenvalues if $3N$ Cartesian coordinates are used. The six zero eigenvalues correspond to the translational and rotational degrees of freedom of the molecule (thus these six zero eigenvalues are not obtained when internal coordinates are used). At a maximum point all eigenvalues are negative and at a saddle point one or more eigenvalues are negative. We will consider the uses of the eigenvalue and eigenvector information in Sections 5.8 and 5.9.

5.7.2 Convergence Criteria

In contrast to the simple analytical functions that we have used to illustrate the operation of the various minimisation methods, in 'real' molecular modelling applications it is rarely possible to identify the 'exact' location of minima and saddle points. We can only ever hope to find an approximation to the true minimum or saddle point. Unless instructed otherwise, most minimisation methods would keep going forever, moving ever closer to the minimum. It is therefore necessary to have some means to decide when the minimisation calculation is sufficiently close to the minimum and so can be terminated. Any calculation is of course limited by the precision with which numbers can be stored on the computer, but in most instances it is usual to stop well before this limit is reached. A simple strategy is to monitor the energy from one iteration to the next and to stop when the difference in energy between successive steps falls below a specified threshold. An alternative is to monitor the change in coordinates and to stop when the difference between successive configurations is sufficiently small. A third method is to calculate the root-mean-square gradient. This is obtained by adding the squares of the gradients of the energy with respect to the coordinates, dividing by the number of coordinates and taking the square root:

$$\mathrm{RMS} = \sqrt{\frac{\mathbf{g}^{\mathrm{T}}\mathbf{g}}{3N}} \tag{5.25}$$

It is also useful to monitor the maximum value of the gradient to ensure that the minimisation has properly relaxed all the degrees of freedom and has not left a large amount of strain in one or two coordinates.

5.8 Applications of Energy Minimisation

Energy minimisation is very widely used in molecular modelling and is an integral part of techniques such as conformational search procedures (Chapter 9). Energy minimisation is also used to prepare a system for other types of calculation. For example, energy minimisation may be used prior to a molecular dynamics or Monte Carlo simulation in order to relieve any unfavourable interactions in the initial configuration of the system. This is especially recommended for simulations of complex systems such as macromolecules or large molecular assemblies. In the following sections we will discuss some techniques that are specifically associated with energy minimisation methods.

5.8.1 Normal Mode Analysis

The molecular mechanics or quantum mechanics energy at an energy minimum corresponds to a hypothetical, motionless state at 0 K. Experimental measurements are made on molecules at a finite temperature when the molecules undergo translational, rotational and vibration motion. To compare the theoretical and experimental results it is

necessary to make appropriate corrections to allow for these motions. These corrections are calculated using standard statistical mechanics formulae. The internal energy $U(T)$ at a temperature T is given by:

$$U(T) = U_{\text{trans}}(T) + U_{\text{rot}}(T) + U_{\text{vib}}(T) + U_{\text{vib}}(0) \tag{5.26}$$

If all translational and rotational modes are fully accessible in accordance with the equipartition theorem, then $U_{\text{trans}}(T)$ and $U_{\text{rot}}(T)$ are both equal to $\frac{3}{2}k_{\text{B}}T$ per molecule (except that $U_{\text{rot}}(T)$ equals $k_{\text{B}}T$ for a linear molecule); k_{B} is Boltzmann's constant. However, the vibrational energy levels are often only partially excited at room temperature. The vibrational contribution to the internal energy at a temperature T thus requires knowledge of the actual vibrational frequencies. The vibrational contribution equals the difference in the vibrational enthalpy at the temperature T and at 0 K and is given by:

$$U_{\text{vib}}(T) = \sum_{i=1}^{N_{\text{nm}}} \left(\frac{h\nu_i}{2} + \frac{h\nu_i}{\exp[h\nu_i/k_{\text{B}}T] + 1} \right) \tag{5 27}$$

N_{nm} is the number of *normal vibrational modes* for the system. Even the zero-point energy ($U_{\text{vib}}(0)$, obtained by summing $\frac{1}{2}h\nu_i$ for each normal mode) can be quite substantial, amounting to about 100 kcal/mol for a six-carbon alkane. Other thermodynamic quantities such as entropies and free energies may also be calculated from the vibrational frequencies using the relevant statistical mechanics expressions.

Normal modes are useful because they correspond to collective motions of the atoms in a coupled system that can be individually excited. The three normal modes of water are schematically illustrated in Figure 5.15; a non-linear molecule with N atoms has $3N - 6$ normal modes. The frequencies of the normal modes together with the displacements of the individual atoms may be calculated from a molecular mechanics force field or from the wavefunction using the Hessian matrix of second derivatives ($\mathscr{V}''$). Of course, if we have used an appropriate minimisation algorithm then we already know the Hessian. The Hessian must first be converted to the equivalent force-constant matrix in *mass-weighted coordinates* (**F**), as follows.

$$\mathbf{F} = \mathbf{M}^{-1/2}\mathscr{V}''\mathbf{M}^{-1/2} \tag{5.28}$$

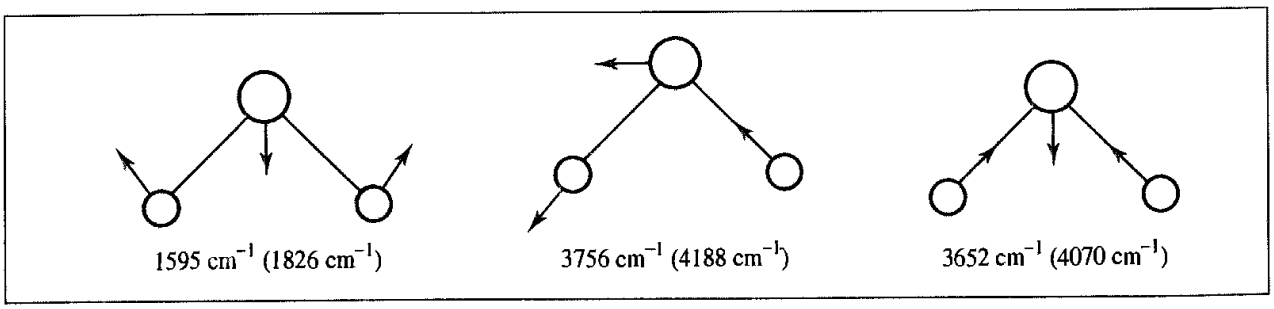

Fig 5.15 Normal modes of water Experimental and (calculated) frequencies are shown. Theoretical frequencies calculated using a 6-31G basis set*

$\mathbf{M}$ is a diagonal matrix of dimension $3N \times 3N$, containing the atomic masses. All elements of $\mathbf{M}$ are zero except those on the diagonal; $M_{1,1} = m_1$, $M_{2,2} = m_1$, $M_{3,3} = m_1$, $M_{4,4} = m_2, \ldots, M_{3N-2,3N-2} = m_N$, $M_{3N-1,3N-1} = m_N$, $M_{3N,3N} = m_N$. Each non-zero element of $\mathbf{M}^{-1/2}$ is thus the inverse square root of the mass of the appropriate atom. The masses of the atoms must be taken into account because a force of a given magnitude will have a different effect upon a larger mass than a smaller one. For example, the force constant for a bond to a deuterium atom is, to a good approximation, the same as to a proton, yet the different mass of the deuteron gives a different motion and a different zero-point energy. The use of mass-weighted coordinates takes care of these problems.

We next solve the secular equation $|\mathbf{F} - \mathbf{I}| = 0$ to obtain the eigenvalues and eigenvectors of the matrix $\mathbf{F}$. This step is usually performed using matrix diagonalisation, as outlined in Section 1.10.3. If the Hessian is defined in terms of Cartesian coordinates then six of these eigenvalues will be zero as they correspond to translational and rotational motion of the entire system. The frequency of each normal mode is then calculated from the eigenvalues using the relationship:

$$\nu_i = \frac{\sqrt{\lambda_i}}{2\pi} \tag{5.29}$$

As a simple example of a normal mode calculation consider the linear triatomic system in Figure 5.16. We shall just consider motion along the long axis of the molecule. The displacements of the atoms from their equilibrium positions along this axis are denoted by ξ_i. It is assumed that the displacements are small compared with the equilibrium values l_0 and the system obeys Hooke's law with bond force constants k. The potential energy is given by:

$$\mathscr{V} = \tfrac{1}{2}k(\xi_1 - \xi_2)^2 + \tfrac{1}{2}k(\xi_2 - \xi_3)^2 \tag{5.30}$$

We next calculate the first and then the second derivatives of the potential energy with respect to the three coordinates ξ_1, ξ_2 and ξ_3:

$$\frac{\partial \mathscr{V}}{\partial \xi_1} = k(\xi_1 - \xi_2); \qquad \frac{\partial \mathscr{V}}{\partial \xi_2} = -k(\xi_1 - \xi_2) + k(\xi_2 - \xi_3); \qquad \frac{\partial \mathscr{V}}{\partial \xi_3} = -k(\xi_2 - \xi_3) \tag{5.31}$$

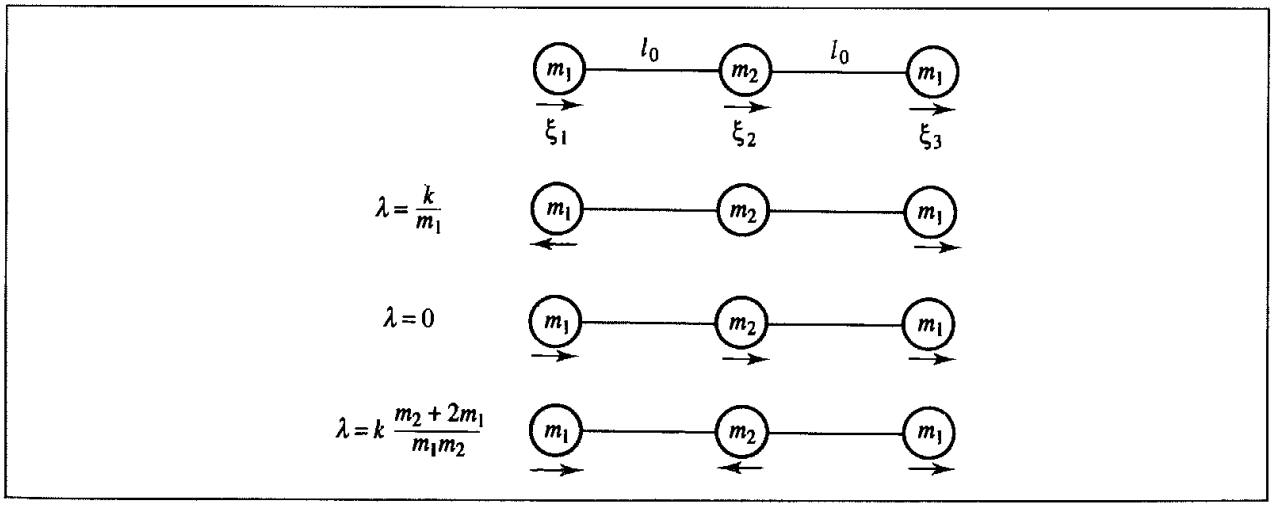

Fig 5.16· Linear three-atom system with results of normal mode calculation

The second derivatives are conveniently represented as a 3×3 matrix:

$$\begin{vmatrix} k & -k & 0 \\ -k & 2k & -k \\ 0 & -k & k \end{vmatrix} \tag{5.32}$$

The mass-weighted matrix is

$$\begin{vmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{vmatrix} \tag{5.33}$$

The secular equation to be solved is thus:

$$\begin{vmatrix} \frac{k}{m_1} - \lambda & -\frac{k}{\sqrt{m_1}\sqrt{m_2}} & 0 \\ -\frac{k}{\sqrt{m_1}\sqrt{m_2}} & \frac{2k}{m_2} - \lambda & -\frac{k}{\sqrt{m_1}\sqrt{m_2}} \\ 0 & -\frac{k}{\sqrt{m_1}\sqrt{m_2}} & \frac{k}{m_1} - \lambda \end{vmatrix} = 0 \tag{5.34}$$

This determinant leads to a cubic in λ which has three roots (λ_k), each corresponding to a different mode of motion:

$$\lambda = \frac{k}{m_1}, \qquad \lambda = 0, \qquad \lambda = k\frac{m_2 + 2m_1}{m_1 m_2} \tag{5.35}$$

The corresponding frequencies can be obtained from Equation (5.29). The amplitudes (A) of each normal mode are given by the eigenvector solutions of the secular equation $\mathbf{FA} = \lambda\mathbf{A}$. If A_1, A_2 and A_3 are the amplitudes of each atom then the amplitudes obtained for each eigenvalue are:

$$\lambda = \frac{k}{m_1}\cdot \qquad A_1 = -A_3; \qquad A_2 = 0 \tag{5.36}$$

$$\lambda = 0: \qquad A_1 = A_3; \qquad A_2 = \sqrt{\frac{m_2}{m_1}}A_1 \tag{5.37}$$

$$\lambda = k\frac{m_2 + 2m_1}{m_1 m_2}\cdot \qquad A_1 = A_3; \qquad A_2 = -2\sqrt{\frac{m_1}{m_2}}A_1 \tag{5 38}$$

These normal modes are schematically illustrated in Figure 5.16. They correspond to a symmetric stretch, a translation and an asymmetric stretch respectively.

We have already seen how the results of normal mode calculations can be used to calculate thermodynamic quantities. The frequencies themselves can also be compared with the results of spectroscopic experiments, information which can be used in the parametrisation of a force field. For example, the experimental frequencies for the normal modes of water are shown in Figure 5.15, together with the frequencies determined using a 6-31G* *ab initio* calculation. The calculated values clearly deviate from those obtained experimentally, but the ratio of the experimental and theoretical frequencies is

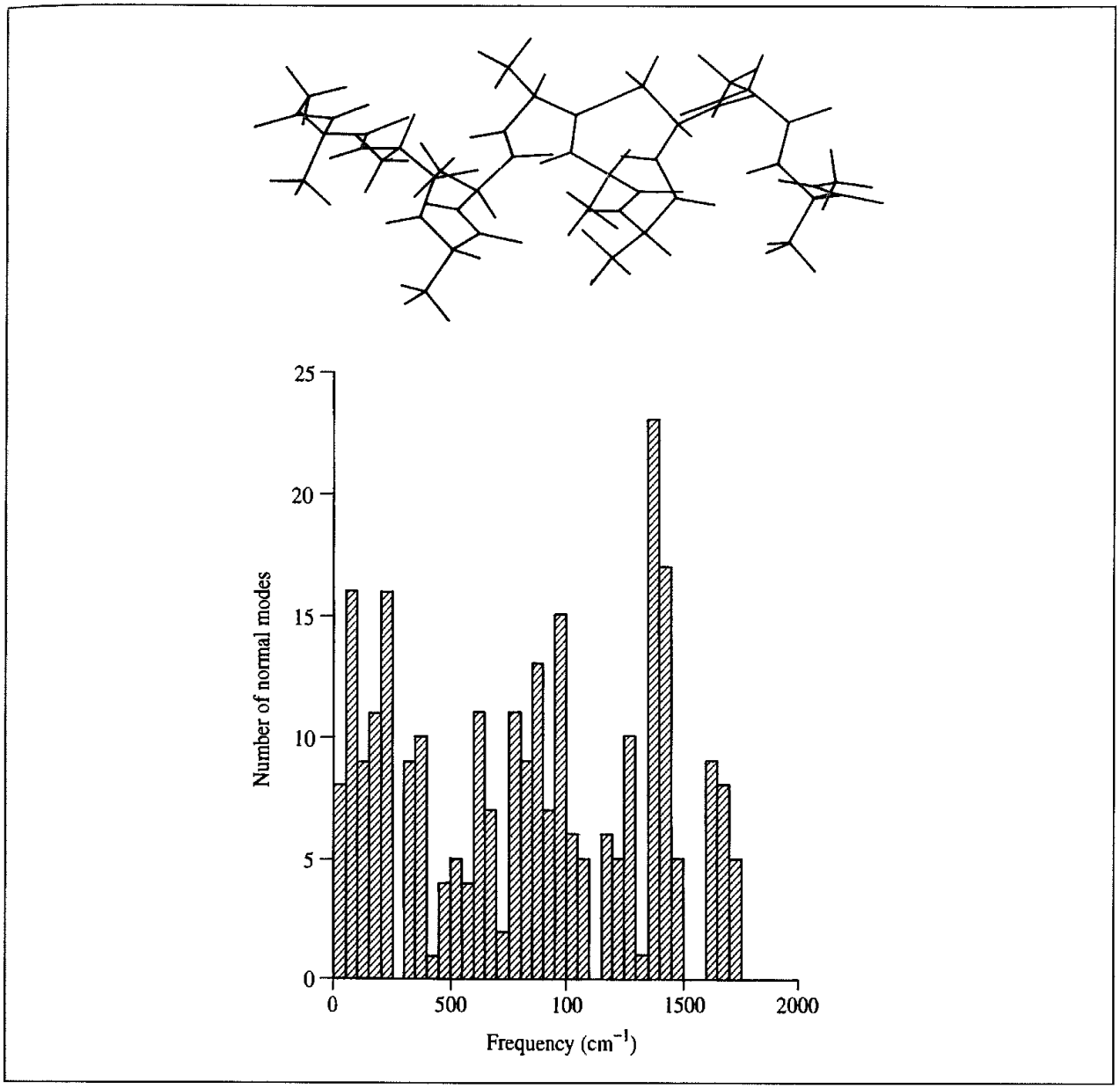

Fig. 5.17. Histogram of the normal modes calculated for a polyalanine polypeptide in an α-helical conformation. The height of each bar indicates the number of normal modes in each 50 cm^{-1} section.

remarkably consistent (at about 1.1). Such empirical scaling factors have been derived which enable frequencies obtained using a given level of theory to be converted to values for experiment or a higher level of theory [Pople *et al.* 1993]. The normal modes of much larger molecules can be calculated using molecular mechanics. For example, the vibrations of a helical polypeptide constructed from a sequence of ten alanine residues (112 atoms) are shown in Figure 5.17. In such cases it is usually the low-frequency vibrations that are of most interest as these correspond to the large-scale conformational motions of the molecule. The results of such analyses can be compared with molecular dynamics simulations from which vibrational contributions can also be extracted [Brooks and Karplus 1983].

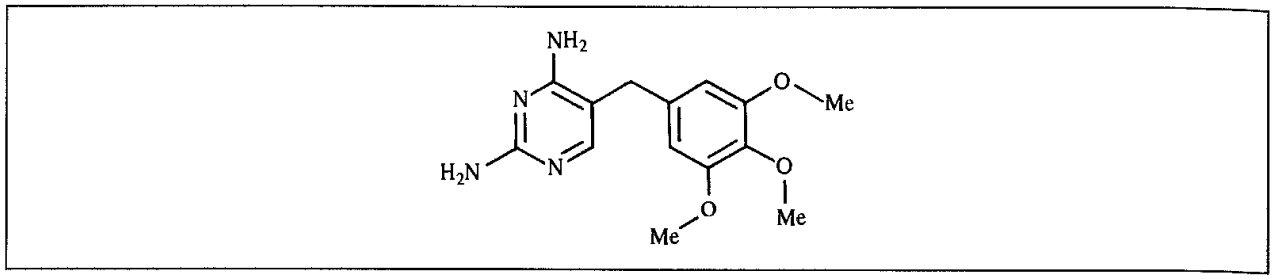

Fig. 5.18: Trimethoprim.

A normal mode calculation is based upon the assumption that the energy surface is quadratic in the vicinity of the energy minimum (the harmonic approximation). Deviations from the harmonic model can require corrections to calculated thermodynamic properties. One way to estimate anharmonic corrections is to calculate a force constant matrix using the atomic motions obtained from a molecular dynamics simulation; such simulations are not restricted to movements on a harmonic energy surface. The eigenvalues and eigenvectors are then calculated for this quasi-harmonic force-constant matrix in the normal way, giving a model which implicitly incorporates the anharmonic effects.

The harmonic approximation to the energy surface is found to be appropriate for well-defined energy minima such as the intramolecular degrees of freedom of small molecules and for some small intermolecular complexes. For larger systems such as liquids and large, 'floppy' molecules, the harmonic approximation breaks down. Such systems also have an extraordinarily large number of 'minima' on the energy surface. In such cases it is not possible to calculate accurately thermodynamic properties using energy minimisation and normal mode calculations. Rather, molecular dynamics or Monte Carlo simulations must be used to sample the energy surface from which properties can be derived, as we will discuss in Chapters 6–8.

5.8.2 The Study of Intermolecular Processes

One example of the use of minimisation methods and normal-mode analysis is the study by Hagler and co-workers of the binding of the antibacterial drug trimethoprim (Figure 5.18) to the enzyme dihydrofolate reductase (DHFR) [Dauber-Osguthorpe *et al.* 1988; Fisher *et al.* 1991]. DHFR catalyses the reduction of folic acid and dihydrofolic acid to tetrahydrofolic acid (Figure 5.19) and plays a vital metabolic role in the biosynthesis of nucleic acids in bacteria, protozoa, plants and animals. Trimethoprim exploits the structural differences between bacterial and vertebrate DHFR, binding much more strongly to the former, and is clinically used as an antibacterial agent. Inhibitors of human DHFR are used in cancer therapy. Hagler and colleagues applied energy minimisation to an isolated trimethoprim molecule, to the crystal structure of trimethoprim, to trimethoprim in the presence of water molecules, and to trimethoprim in intermolecular complexes with DHFR from both bacterial and vertebrate sources. An important observation was that the conformation of the trimethoprim, when bound to the enzyme, was significantly different from that obtained for the isolated molecule. This reinforces the view that the use of

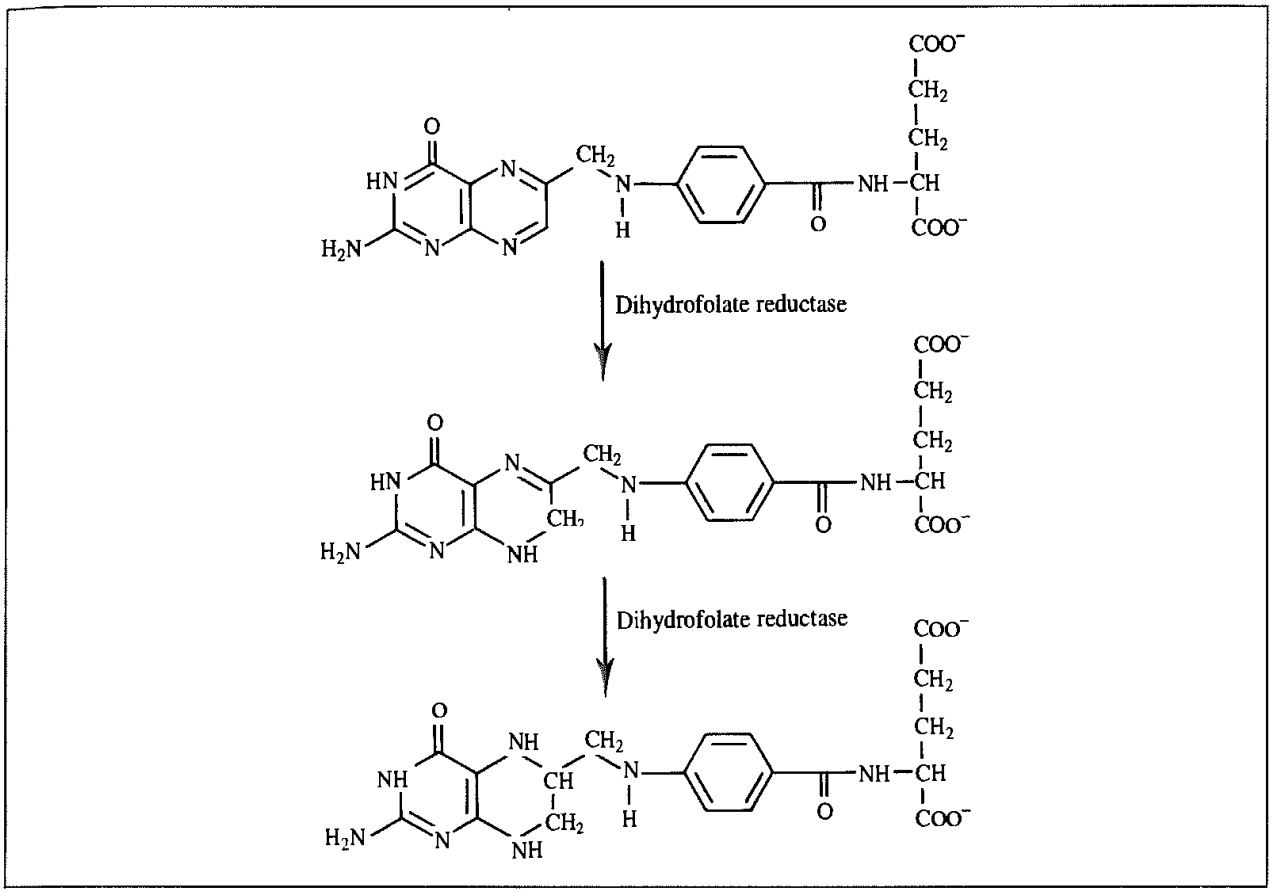

Fig 5.19: DHFR catalyses the reduction of folic acid to tetrahydrofolic acid.

structures obtained from energy minimisation calculations on isolated molecules can lead to misleading conclusions. Intermolecular interactions with the receptor enable the ligand to adopt a conformation whose intramolecular energy is significantly higher than any of its minimum energy structures.

A normal mode analysis on the isolated and bound trimethoprim molecules enabled an estimate to be made of the entropic contribution to binding. Low-frequency modes for the isolated ligand were found to be shifted to higher frequencies for the ligand in the enzyme complex, reflecting a restriction of the motion of the ligand by the protein. This entropic contribution to the free energy of binding was predicted to be quite significant, indicating that conclusions based solely upon energies may be misleading.

5.9 Determination of Transition Structures and Reaction Pathways

Chemists are interested not only in the thermodynamics of a process (the relative stability of the various species) but also in its kinetics (the rate of conversion from one structure to another). Knowledge of the minimum points on an energy surface enables thermodynamic data to be interpreted, but for the kinetics it is necessary to investigate the nature of the

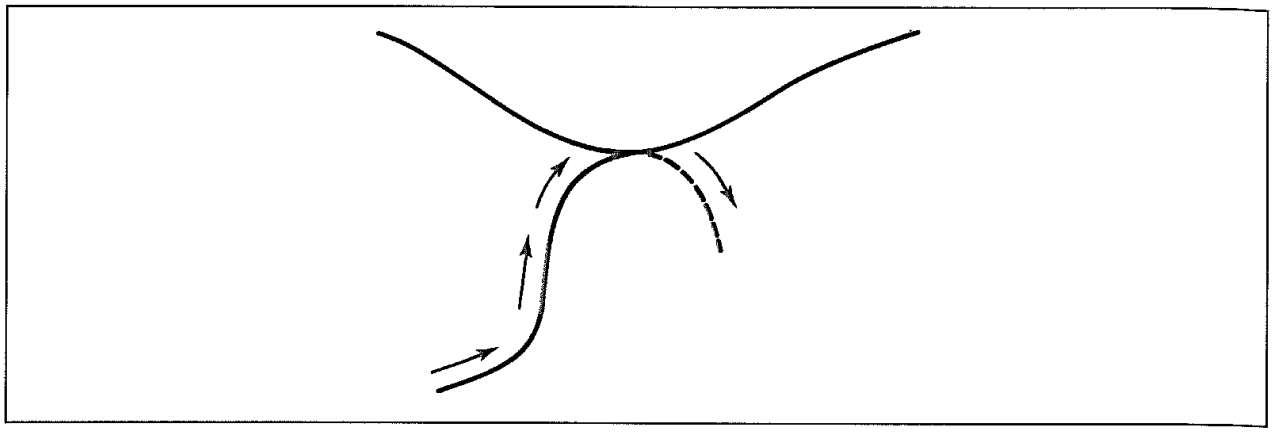

Fig. 5.20: The lowest-energy path from one minimum to another passes through a saddle point

energy surface away from the minimum points. In particular, we would like to know how the system changes from one minimum to another. What changes in geometry are involved, and how does the energy vary during the transition? The minimum points on the energy surface may be the reactants and products of a chemical reaction, two conformations of a molecule, or two molecules that associate to form a non-covalently bound bimolecular complex. We shall use the term 'reaction pathway' to describe the path between two minima, but our use of the word 'reaction' does not necessarily mean that bond making and/or breaking is involved. Many methods have been proposed for finding transition structures and elucidating reaction pathways. We do not have space to cover all of the methods, and so we shall restrict our discussion to some of the more common approaches.

As a system moves from one minimum to another, the energy increases to a maximum at the transition structure and then falls. At a saddle point the first derivatives of the potential function with respect to the coordinates are all zero (just as they are at a minimum point). The number of negative eigenvalues in the Hessian matrix is used to distinguish different types of saddle point; an *n*th-order transition or saddle point has *n* negative eigenvalues. We are usually most interested in first-order saddle points, where the energy passes through a maximum for movement along the pathway that connects the two minima, but is a minimum for displacements in all other directions perpendicular to the path. This is shown schematically for a two-dimensional energy surface in Figure 5.20.

These negative eigenvalues of the Hessian matrix are often referred to as the 'imaginary' frequencies for motion of the system over the saddle point. We can illustrate this concept using the gas-phase S_N2 reaction between Cl^- and CH_3Cl. As the chloride ion approaches the methyl chloride along the line of the C–Cl bond the energy passes through an ion-dipole complex which is at an energy minimum. The energy then rises to a maximum at the pentagonal transition state. The energy profile is drawn in Figure 5 21. The geometries of the minimum and the pentagonal transition state, as determined by an *ab initio* HF/SCF calculation with the 6-31G* basis set are shown in Figure 5.22. The lowest-frequency eigenvalues and a representation of the corresponding eigenvectors for the two geometries are also given in Figure 5 22. There are three frequencies in the ion–dipole minimum that are of particularly low energy; two of these correspond to degenerate 'wagging' motions of the

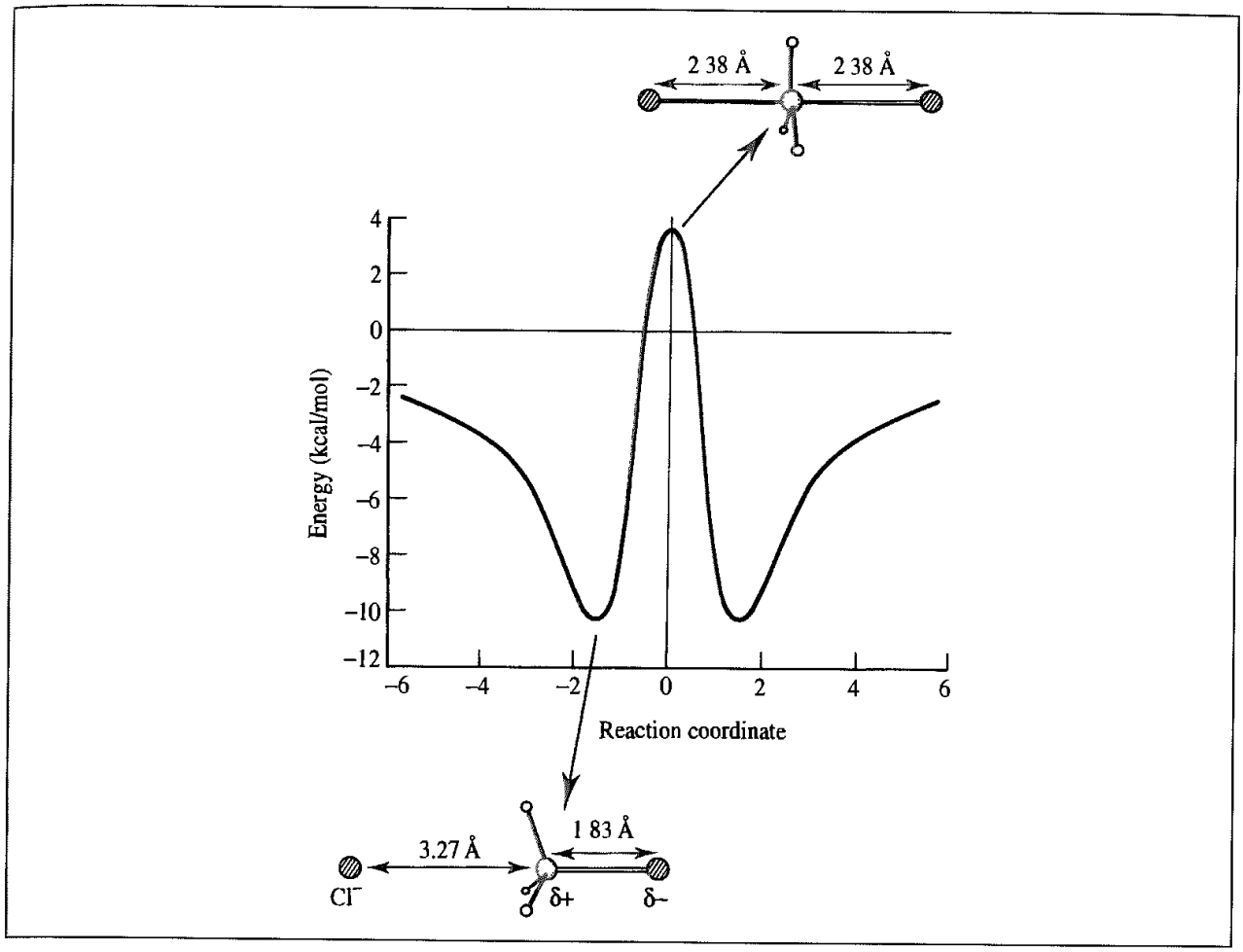

Fig 5 21. The energy profile for the gas-phase Cl^- + MeCl reaction (Adapted in part from Chandrasekhar J, S F Smith and W L Jorgensen 1985 Theoretical Examination of the S_N2 Reaction Involving Chloride Ion and Methyl Chloride in the Gas Phase and Aqueous Solution Journal of the American Chemical Society ***107*** *154–163.)*

system (at 71.3 cm^{-1}). The vibration at 101.0 cm^{-1} is the normal mode that corresponds to motion towards the transition state. At the saddle point there is a single negative eigenvalue (with an imaginary 'frequency' of −415.0 cm^{-1}) that corresponds to vibration along the Cl–C–Cl axis (i.e. motion along the reaction pathway). The other normal modes at the saddle point all have positive frequencies; the two lowest (at 204.2 cm^{-1}) correspond to wagging motions perpendicular to the Cl–C–Cl axis and the third is a symmetric stretch of the two chlorine atoms along the symmetry axis.

It is important to distinguish the transition *structure* from the transition *state*. The transition structure is the point of highest potential energy along the pathway. By contrast, the transition state is the geometry at the peak in the free energy profile. In many cases the geometry at the transition state is very similar to that of the transition structure. However, the transition state may be different as the free energy of activation includes contributions from sources other than just the potential energy. If the transition state geometry is temperature-dependent then entropic factors may be important. An example is the following radical

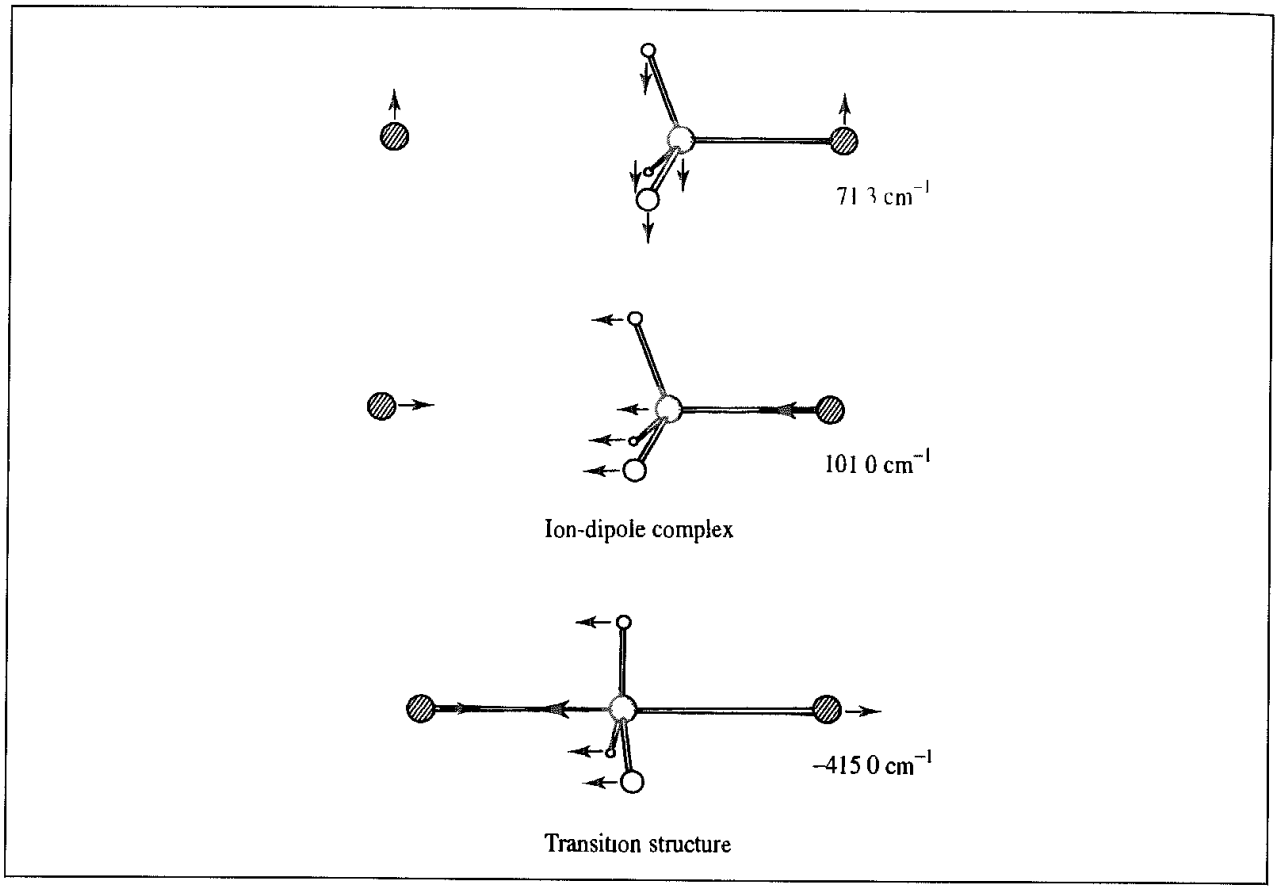

Fig 5 22 Schematic representation of some of the lower frequencies in the ion–dipole complex for the Cl^- + MeCl reaction and the imaginary frequency of the transition structure, calculated using a 6-31G basis set.*

reaction:

$$H^{\bullet} + CH_3CH_2^{\bullet} \rightarrow H_2 + CH_2{=}CH_2$$

The calculated geometry of the transition structure resembles the ethyl radical (Figure 5.23) [Doubleday *et al.* 1985]. The entropy change for this reaction is negative and so, as the temperature is increased, the maximum in the free energy profile shifts more towards the products, in the direction of lower entropy.

Methods for finding transition structures and reaction pathways are often closely related. Thus, some methods for finding the reaction pathway start from the transition structure and move down towards a minimum. Such methods must be supplied with the transition structure geometry as the starting point. Conversely, some methods for locating transition structures do so by searching along the reaction pathway, or an approximation to it. Yet other methods require neither the transition structure nor the pathway, but can determine both simultaneously from the two minima. In general, it is more difficult to locate transition structures and determine reaction pathways than to find minimum points. It is therefore crucial to check that the Hessian matrix at any proposed saddle point has the required single negative eigenvalue. Methods for locating saddle points are usually most effective when given as input a geometry that is as close as possible to the transition structure. It

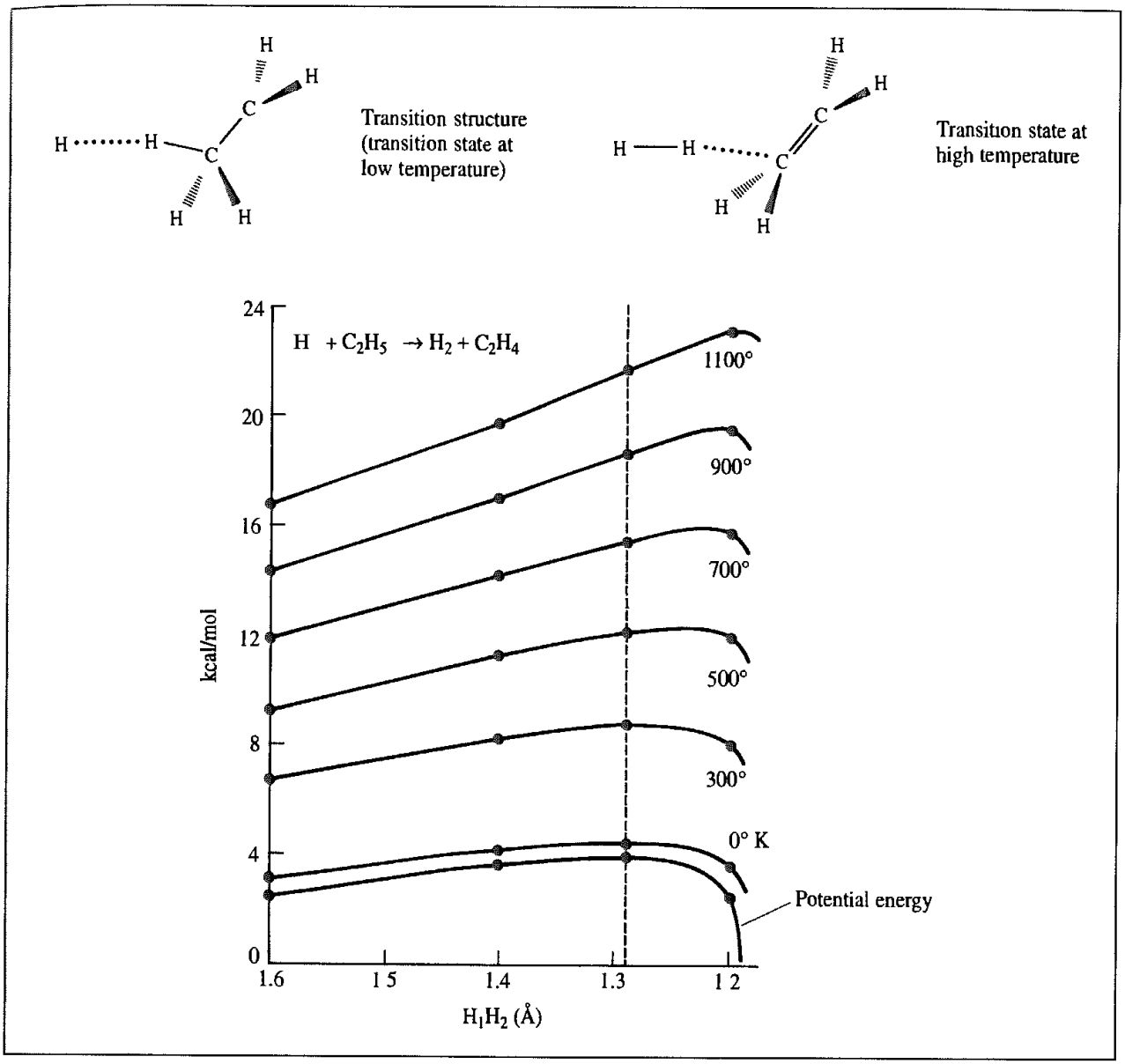

Fig 5 23 The transition structure for the $H^{\cdot} + CH_3CH_2^{\cdot} \rightarrow H_2 + CH_2{=}CH_2$ *reaction. At low temperature the transition structure corresponds to the transition state (maximum of free energy) At high temperature the transition state moves closer to the products, as can be seen from the graph. (Redrawn from Doubleday C, J McIver, M Page and T Zielinski 1985. Temperature Dependence of the Transition-State Structure for the Disproportionation of Hydrogen Atom with Ethyl Radical.* Journal of the American Chemical Society ***107***:*5800–5801*)

can also be helpful to examine the atomic displacements that correspond to the negative eigenvector, to ensure that it corresponds to the correct motion over the saddle point as for the $Cl^- + CH_3Cl$ reaction.

As one approaches the saddle point from a minimum, the Hessian matrix will change from having all positive eigenvalues to including one negative value. The *quadratic region* of a saddle point is that portion of the energy surface surrounding the point where the Hessian contains one negative eigenvalue. Similarly the quadratic region of a minimum is the

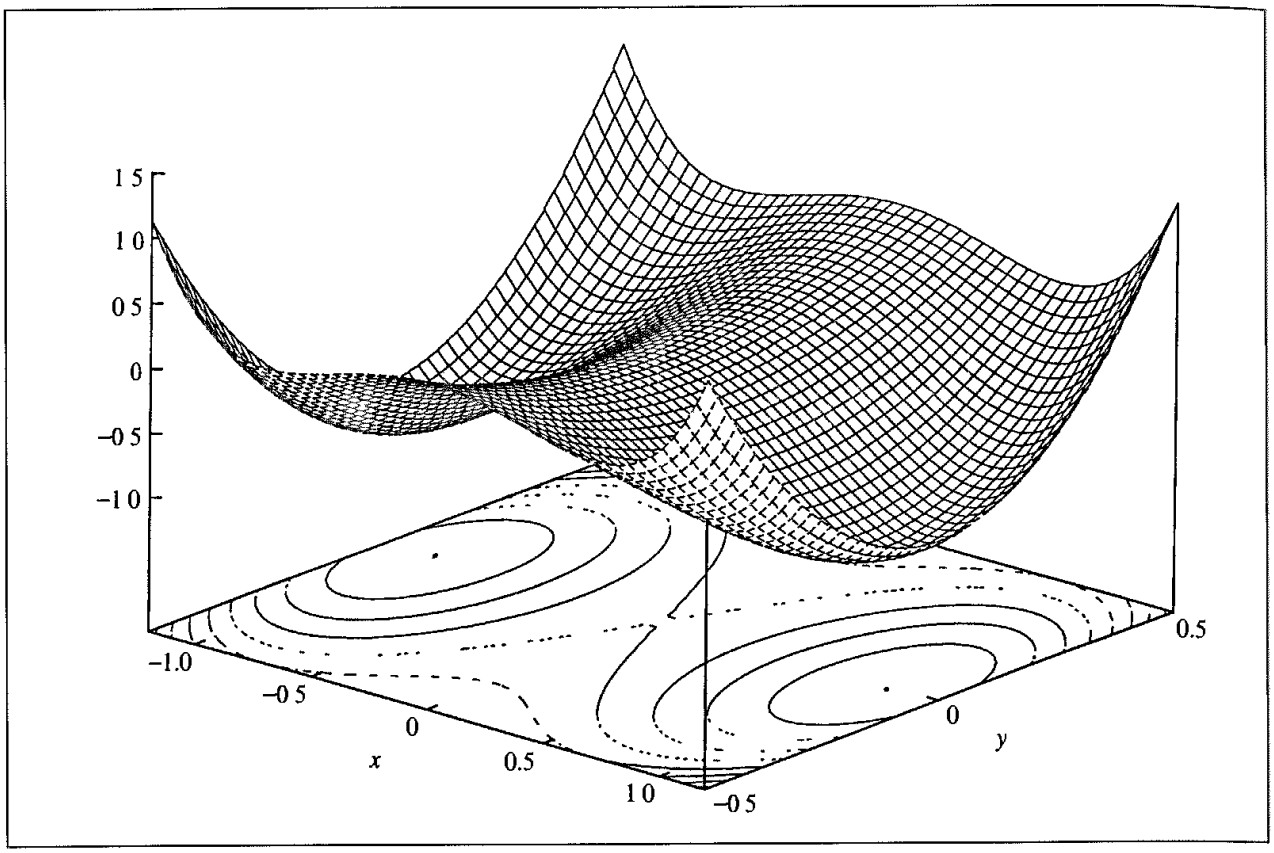

Fig. 5 24. The function $f(x,y) = x^4 + 4x^2y^2 - 2x^2 + 2y^2$ has a saddle point at $(0,0)$ and minima at $(1,0)$ and $(-1,0)$

region where all eigenvalues are positive and the Hessian is positive definite. Some algorithms for finding saddle points require a starting geometry within the quadratic region. We can illustrate the concept of a quadratic region by considering the function $f(x,y) = x^4 + 4x^2y^2 - 2x^2 + 2y^2$, which is drawn in Figure 5.24. This function has two minima at $(1,0)$ and $(-1,0)$ and one saddle point at $(0,0)$. In this case it is possible to derive and characterise the stationary points analytically. The Hessian matrix of second derivatives for this function is:

$$\begin{pmatrix} 12x^2 + 8y^2 - 4 & 16xy \\ 16xy & 8x^2 + 4 \end{pmatrix} \tag{5.39}$$

At the point $(1,0)$ the Hessian matrix is thus

$$\begin{pmatrix} 8 & 0 \\ 0 & 4 \end{pmatrix} \tag{5.40}$$

The eigenvalues of this matrix are obtained by setting the secular determinant to zero:

$$\begin{vmatrix} 8-\lambda & 0 \\ 0 & 4-\lambda \end{vmatrix} = 0 \tag{5 41}$$

The eigenvalues are $\lambda = 4$ and $\lambda = 8$. Thus both eigenvalues are positive and the point is a minimum. At the point $(0,0)$ the Hessian matrix is

$$\begin{pmatrix} -4 & 0 \\ 0 & 4 \end{pmatrix} \tag{5.42}$$

with one negative and one positive eigenvalue (-4 and $+4$). The normalised eigenvectors corresponding to these eigenvalues are $(0,1)$ for the eigenvalue $\lambda = 4$ and $(1,0)$ for the eigenvalue $\lambda = -4$. These eigenvectors indicate the directions in which the gradient of the function changes sign. Thus along the line $x = 0$ the function passes through a minimum, as can be seen from Figure 5.24. By contrast, if one progresses from $(-1,0)$ to $(1,0)$ through the origin then the function passes through a maximum. As one progresses through a transition structure the eigenvector of the negative eigenvalue corresponds to the concerted motions of the atoms that give rise to motion through the saddle point. If we move along the x axis from the minimum at $(1,0)$ to the saddle point at the origin, both eigenvalues will be positive so long as $12x^2 + 8y^2 - 4 > 0$. Thus, so long as x is larger than $1/\sqrt{3}$ the eigenvalues of the Hessian matrix will be positive. When x becomes smaller than $1/\sqrt{3}$ there will be one negative and one positive eigenvalue. In this case the quadratic region would correspond to all points where the absolute value of x was less than $1/\sqrt{3}$.

5.9.1 Methods to Locate Saddle Points

In some simple cases such as the chloride/methyl chloride reaction the geometry of the transition structure can be predicted by inspection. In other cases a *grid search* can be used to scan the energy surface in order to locate the approximate position of the transition state. In a grid search, the coordinates are systematically varied to generate a set of structures, for each of which the energy is calculated. It may then be possible to fit an analytical expression to these points, from which the saddle point can be predicted by standard calculus methods. The grid search method is widely used for constructing potential energy surfaces but is restricted to systems with a very small number of atoms or where only a limited number of degrees of freedom are being explored such as the $H + H_2 \rightarrow H_2 + H$ reaction. An advantage of the grid search is that it does provide information about the energy surface away from the pathway, which can be important if one wishes to investigate the dynamics of a reaction and the interconversion of energy between different modes. The grid search method is not the method of choice for all but the smallest systems due to the number of energy evaluations that are required. In any case it does not directly provide the transition structure.

The conversion of one minimum-energy structure into another may sometimes occur primarily along just one or two coordinates. In such cases, an approximation to the reaction pathway can be obtained by gradually changing the coordinate(s), allowing the system to relax at each stage using minimisation while keeping the chosen coordinate(s) fixed. The point of highest energy on the path is an approximation to the saddle point and the structures generated during the course of the calculation can be considered to represent a sequence of points on the interconversion pathway. When such coordinate driving methods are applied to conformational changes that occur primarily via rotation about bonds, the

the Hessian matrix of second derivatives is available then the appropriate direction to take is uphill along the eigenvector of the smallest eigenvalue when all eigenvalues are positive and downhill along the eigenvector corresponding to the negative eigenvalue when within the quadratic region of the saddle point [Baker 1986].

As we have stated frequently, at a saddle point the gradient is zero (as it is for a minimum). It might therefore be imagined that a minimisation algorithm (or some variant) could be used to locate saddle points. Some minimisation algorithms can occasionally incorrectly converge to a saddle point, especially if the starting structure is close to the transition structure A simple example is the Newton-Raphson method, which will converge to a transition structure when giving a starting position that is within the quadratic region. Other minimisation algorithms can also be modified so that they consistently locate saddle points when provided with an initial structure within the quadratic region [Schlegel 1982].

5.9.2 Reaction Path Following

The traditional way to elucidate the reaction path is to move downhill from a saddle point to the two associated minima. There may be many different paths that could be followed from the saddle point to the associated minima. The *intrinsic reaction coordinate* (IRC) is the path that would be followed by a particle moving along the steepest descents path with an infinitely small step from the transition structure down to each minimum when the system is described using mass-weighted coordinates (as in a normal mode calculation) [Fukui 1981]. The initial directions towards each minimum can be obtained directly from the eigenvector that corresponds to the imaginary frequency at the transition structure. A simple steepest descents algorithm with a reasonable step size will usually give a path that oscillates about the true minimum energy path, as illustrated in Figure 5.28. This is perfectly acceptable in a minimisation, where the objective is to locate the minimum as efficiently as possible and where we are not interested in the intermediate structures. To determine the true reaction pathway (or a better approximation to it) it is necessary to 'correct' the path taken by the steepest descents algorithm. These corrective methods are especially useful when the path is curved.

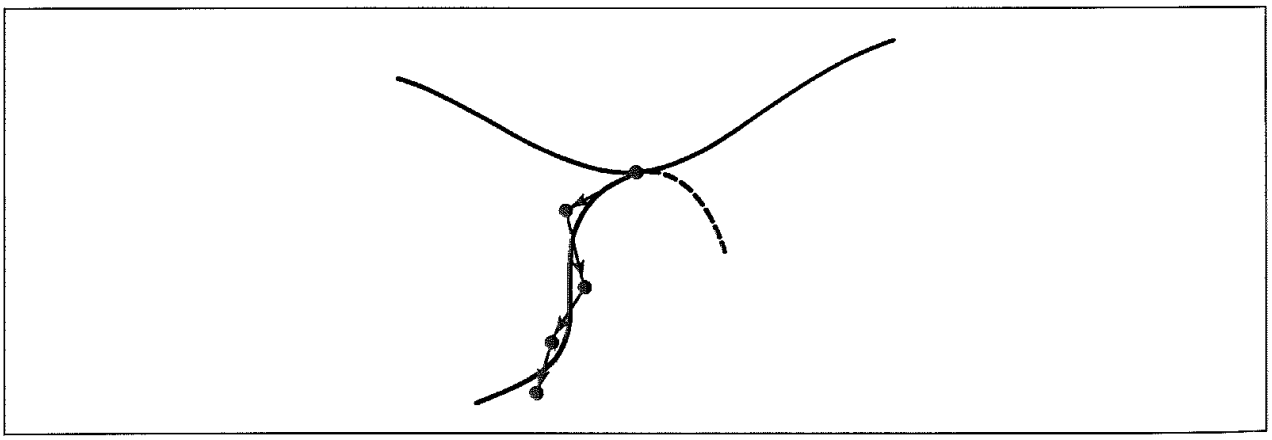

Fig 5 28 A steepest descents minimisation algorithm produces a path that oscillates about the true reaction pathway from the transition structure to a minimum

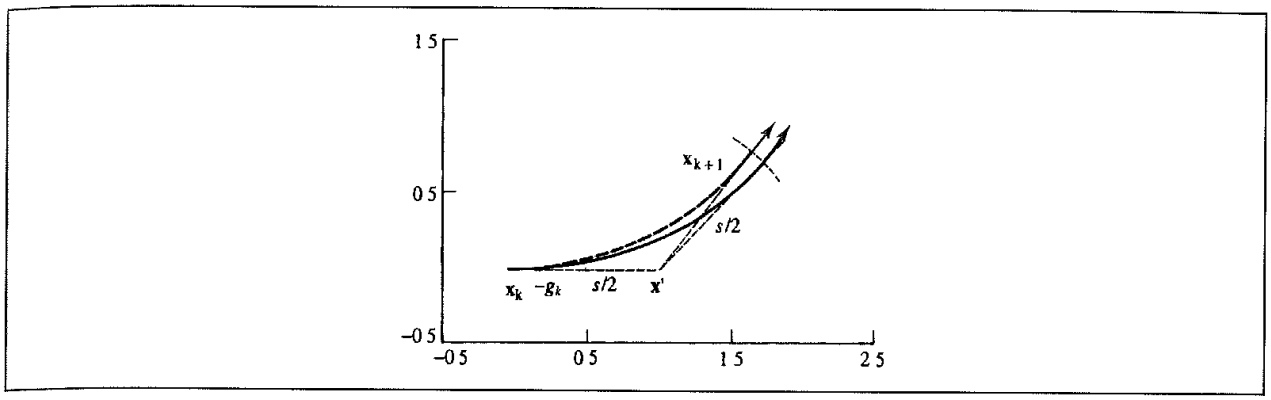

Fig 5 29: Method for correcting the path followed by a steepest descents algorithm to generate the intrinsic reaction coordinate. The solid line shows the real path and the dotted line shows the algorithmic approximation to it (Figure redrawn from Gonzalez C and H B Schlegel 1988 An Improved Algorithm for Reaction Path Following Journal of Chemical Physics ***90*** *2154–2161)*

Many different algorithms have been suggested for determining reaction paths The real challenge is to find an approach that is sufficiently general to work well in many (if not all) situations and with relatively little computational expense. One widely used method was devised by Gonzalez and Schlegel [Gonzalez and Schlegel 1988] and is illustrated in Figure 5 29. First it calculates the gradient at the current point, $\mathbf{x}_k$. A step of length $s/2$ is taken along the direction of this gradient to give a new point ($\mathbf{x}'$). The next point on the reaction path is obtained by minimising the energy subject to the constraint that the distance between $\mathbf{x}'$ and the new point on the reaction path ($\mathbf{x}_{k+1}$) is $s/2$. The reaction path is then approximated by a circle that passes through both $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ and whose tangents at those two points are in the directions of the gradients. A refined version of this path-following algorithm has been incorporated into an efficient combined procedure which can determine reaction paths, minima and transition state geometries [Ayala and Schlegel 1997] without the need for second derivatives to be calculated.

5.9.3 Transition Structures and Reaction Pathways for Large Systems

Most of the algorithms we have discussed so far, with the possible exception of adiabatic mapping, were originally designed to be used with quantum mechanics where relatively small numbers of atoms are involved. It is often difficult to apply these methods to the study of conformational transitions. There are several reasons for this, but one important feature is that it is assumed that there is only one saddle point between the initial and final states. There may be a number of transition structures along the pathway between two conformations of a complex molecule. Here we will discuss two related methods that were originally designed to tackle this problem using molecular mechanics.

In the self-penalty walk (SPW) method of Czerminski and Elber [Czerminski and Elber 1990; Nowak *et al.* 1991] a 'polymer' is constructed that consists of a series of $M+2$ 'monomers'. Each monomer is a complete copy of the actual system and so there are $(M+2)N$ atoms present in the calculation. The two ends of the polymer correspond to the two minima between which we are trying to elucidate the pathway (the 'reactant' and the 'product').

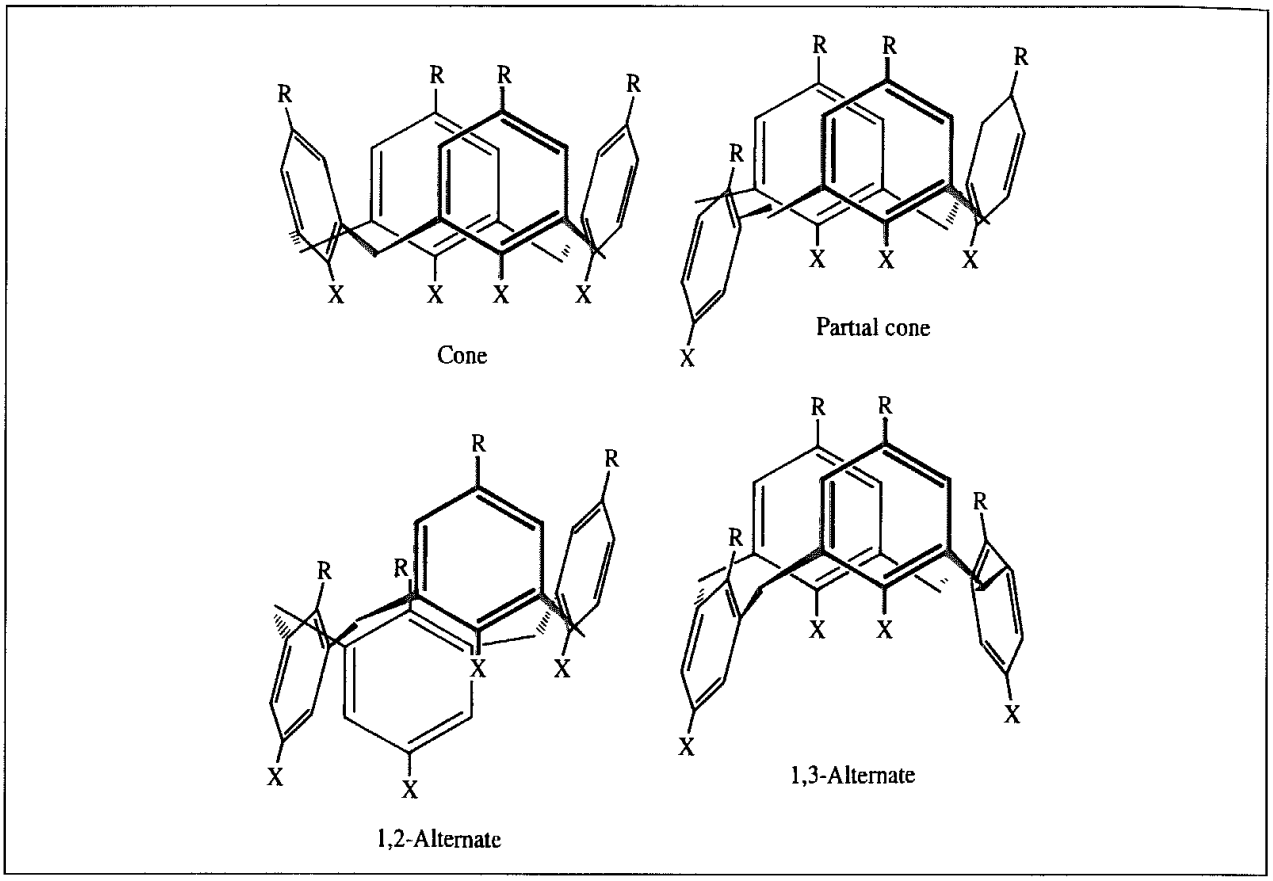

Fig. 5.31 Possible conformations of the calix[4]arene systems. (Figure adapted from Fischer S, P D J Groothenuis, L C Groenen, W P van Hoorn, F C J M van Geggel, D N Reinhoudt and M Karplus 1995 Pathways for Conformational Interconversion of Calix[4]arenes Journal of the American Chemical Society *117.1611–1620)*

involved, giving the energy diagram in Figure 5.32. The predicted activation barrier of 14.5 kcal/mol for the cone → inverted cone transition was in very good agreement with the experimentally determined value of 14.2 kcal/mol. Much of the barrier (9.1 kcal/mol) was due to the need to break two hydrogen bonds; the remainder was due to the need to deform some bond angles such as those of the bridging methylene carbons.

5.9.4 The Transition Structures of Pericyclic Reactions

One of the most celebrated examples of the use of quantum mechanical methods in understanding chemical reactivity is the work of Woodward and Hoffmann [Woodward and Hoffmann 1969] who were able to explain the experimentally observed nature of certain types of concerted reaction. The reactions which they studied include cycloadditions, sigmatropic rearrangements, cheletropic reactions, electrocyclic reactions and the ene reaction (Figure 5.33) and are collectively known as pericyclic reactions. The products obtained from such reactions can be understood in terms of simple mechanistic arguments, but such arguments cannot explain some aspects. In particular, the reactions are often highly stereospecific with the reaction rates and the stereoselectivity changes dramatically with

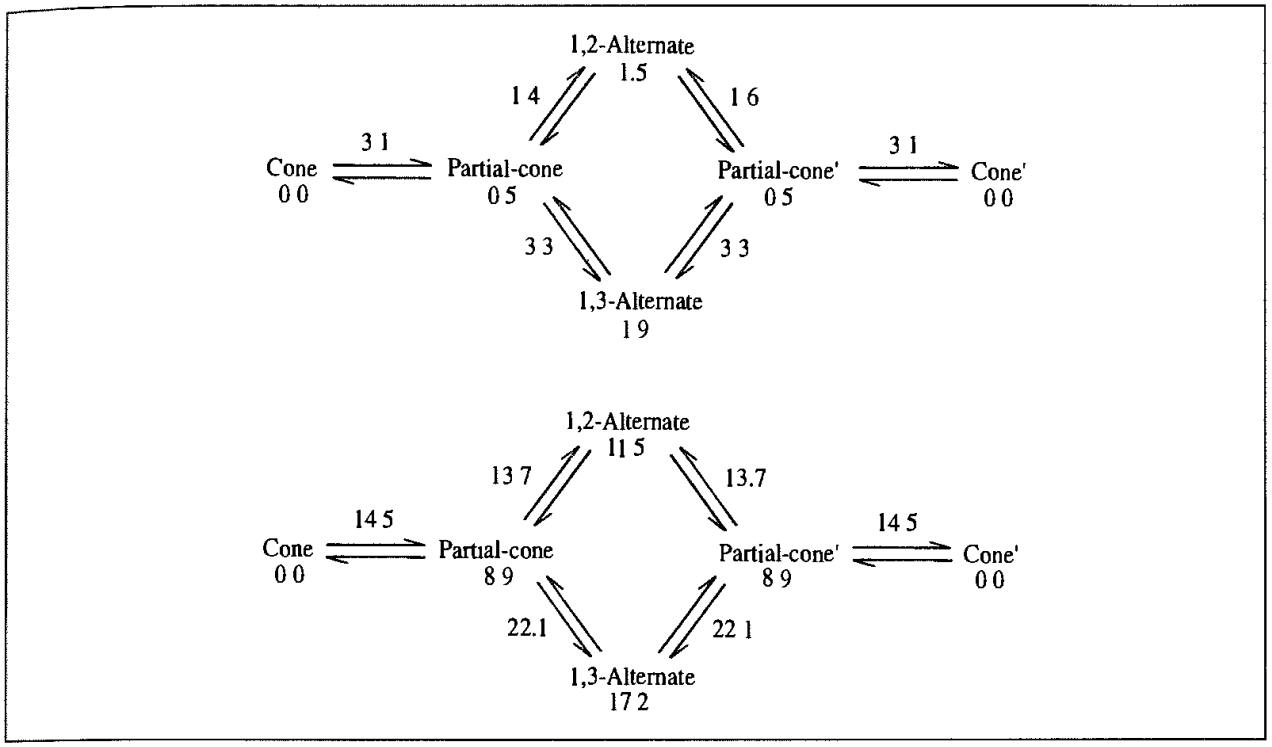

Fig 5 32 Interconversion between various conformations of calix[4]arenes X = H, R = H (top); X = OH, R = H (bottom). Energies in kcal/mol.

the reaction conditions. Woodward and Hoffmann successfully employed molecular orbital theory to rationalise the existing data and their theory has also been very successful in predicting the outcome of similar reactions. The basic principle applied by Woodward and Hoffmann was that of the conservation of orbital symmetry and as a consequence of their work a series of rules (often called the Woodward–Hoffmann rules) were developed. The Woodward–Hoffmann rules apply only to concerted reactions and are based upon the principle that maximum bonding is maintained throughout the course of a reaction. Fukui also discovered the importance of orbital symmetry and suggested that the majority of chemical reactions should take place at the position of, and in the direction of, maximum overlap between the highest occupied molecular orbital (HOMO) of one species and the lowest unoccupied molecular orbital (LUMO) of the other component [Fukui 1971]. These orbitals are collectively known as the *frontier orbitals*.

The HOMO–LUMO interaction depends on various factors, including the geometry of approach (which affects the amount of overlap), the phase relationship of the orbitals and their energy separation. For example, the HOMO and LUMO of ethene are illustrated pictorially in Figure 5.34. The most obvious mode of interaction between the two molecules involves suprafacial attack shown in Figure 5.34 to give cyclobutane. However, the symmetries of the overlapping orbitals must have the same phase for a favourable interaction to occur and this is not possible for ethene unless an energetically unfavourable antarafacial approach is adopted. By contrast, the interaction between ethene and the butadiene does occur in a suprafacial sense with both HOMO/LUMO pairs of orbitals having the appropriate phase relationship (Figure 5.34).

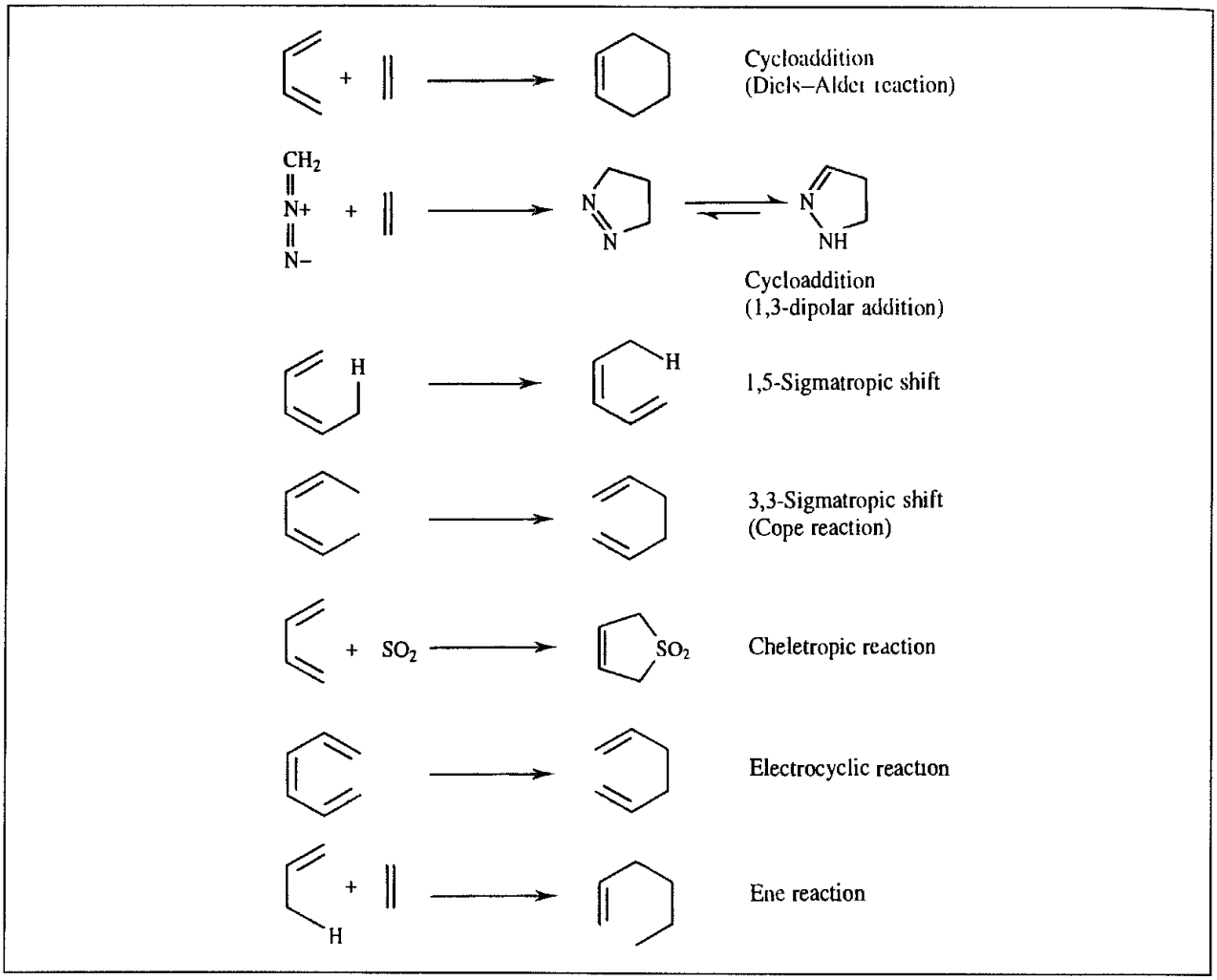

Fig. 5.33: Typical pericyclic reactions.

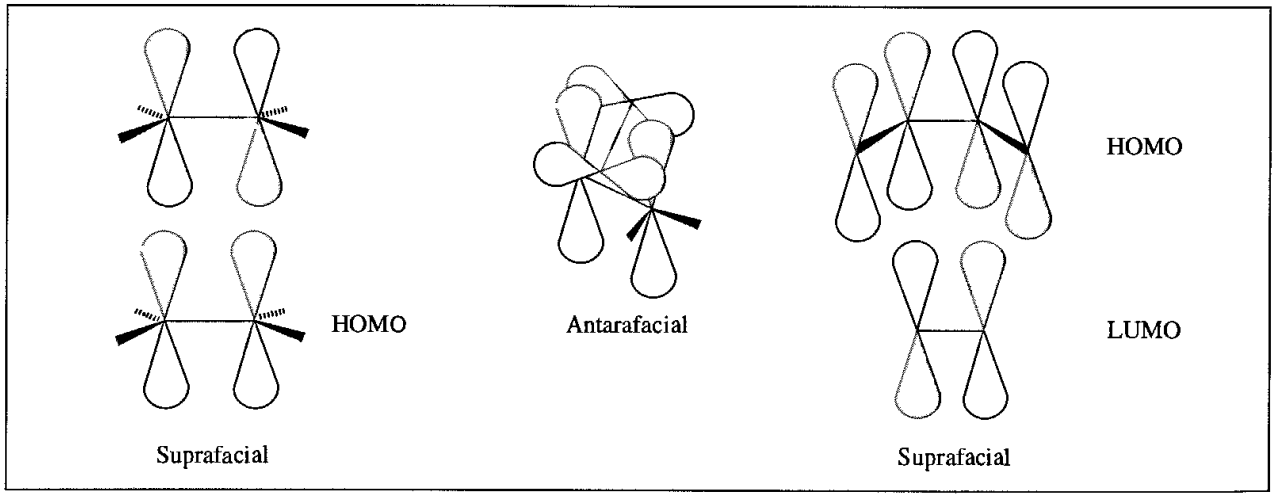

Fig 5 34· Suprafacial attack of one ethene molecule on another (left) is not permitted by the Woodward–Hoffmann rules and the alternative antarafacial mode of attack is sterically unfavourable Suprafacial attack is however permitted for the Diels–Alder reaction between butadiene and ethene (right).

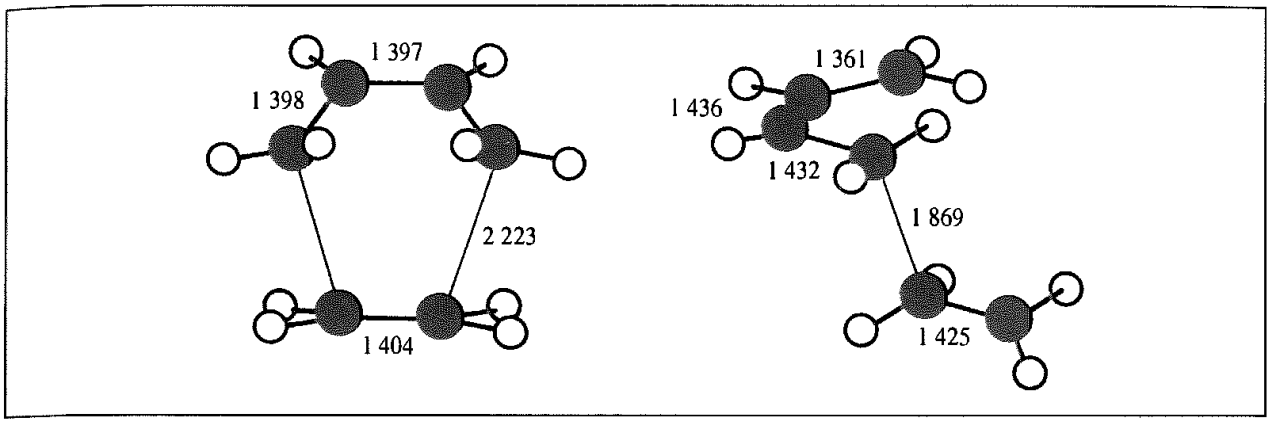

Fig 5 35 Geometry predicted by CASSCF ab initio *calculations of the two possible transition structure geometries for the Diels–Alder reaction between ethene and butadiene (Figure adapted from Houk K N, J González and Y Li 1995 Pericyclic Reaction Transition States· Passions and Punctilios 1935–1995* Accounts of Chemical Research *28 81–90.)*

The Woodward–Hoffmann rules state what the outcome of a pericyclic reaction will be, but they do not define the mechanism by which the reaction occurs. Many theoretical techniques have been applied to the study of these problems over the years [Houk *et al.* 1992] and a passionate debate has ensued on the nature of the transition structures involved in these reactions. The debate has been fuelled by the fact that different theoretical treatments (especially semi-empirical methods) give different results. For example, at one extreme the Diels–Alder reaction between butadiene and ethene would proceed via a two-step mechanism involving a biradial transition structure. At the other extreme the reaction would involve a symmetrical transition state formed in a concerted, synchronous reaction. *Ab initio* calculations at various levels of theory suggest the concerted transition structure. The geometry obtained for the prototypical Diels–Alder reaction between butadiene/ethene using a CASSCF calculation and a 6-31G* basis set is shown in Figure 5.35 [Houk *et al* 1995]. The alternative biradial structure is also shown in Figure 5.35; this is predicted to be 6 kcal/mol higher in energy than the symmetrical transition structure.

5.10 Solid-state Systems: Lattice Statics and Lattice Dynamics

Energy minimisation and normal mode analysis have an important role to play in the study of the solid state. Algorithms similar to those discussed above are employed but an extra feature of such systems, at least when they form a perfect lattice, is that it is can be possible to exploit the space group symmetry of the lattice to speed up the calculations. It is also important to properly take the interactions with atoms in neighbouring cells into account.

The most straightforward type of lattice minimisation is performed at constant volume, where the dimensions of the basic unit cell do not change. A more advanced type of calculation is one performed at constant pressure, in which case there are forces on both the atoms and the unit cell as a whole. The lattice vectors are considered as additional variables along with the atomic coordinates. The laws of elasticity describe the behaviour of a material when

subjected to a *stress* (defined as the force per unit area). One obvious source of stress is any external pressure, but stress may also arise from other sources, especially from interatomic forces within the cell, which give rise to 'internal stress'. The concept of *strain* is also key to this subject; the strain is the fractional change in the dimension (for example, the change per unit length when a steel rod is stretched). In the general case we consider a situation where a point $\mathbf{r}$ in the unstrained material moves to a new point $\mathbf{r}'$ under the effect of some strain:

$$\mathbf{u} = \mathbf{r}' - \mathbf{r} \tag{5.48}$$

If we apply the strain uniformly in one dimension (e.g. the x axis) then the x coordinate of a point that was initially at x will change by an amount proportional to x. This is written:

$$u_x = \varepsilon_{xx} x \tag{5.49}$$

In the general case the constant of proportionality is written as the first derivative:

$$\varepsilon_{xx} = \partial u_x / \partial x \tag{5.50}$$

Deformation in the y and z directions is described in an analogous manner. In order to cater for shear-type strains additional elements are defined.

$$\varepsilon_{xy} = \varepsilon_{yx} = \tfrac{1}{2}(\partial u_y / \partial x + \partial u_x / \partial y) \tag{5.51}$$

These values ε give rise to a *strain tensor* (see Section 4.9.1 for more discussion on tensors), which is symmetric and is often written in the following form:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \frac{1}{2}\varepsilon_6 & \frac{1}{2}\varepsilon_5 \\ \frac{1}{2}\varepsilon_6 & \varepsilon_2 & \frac{1}{2}\varepsilon_4 \\ \frac{1}{2}\varepsilon_5 & \frac{1}{2}\varepsilon_4 & \varepsilon_3 \end{bmatrix} \quad \begin{aligned} &\varepsilon_1 \equiv \varepsilon_{xx}, \quad \varepsilon_2 \equiv \varepsilon_{yy}, \quad \varepsilon_3 \equiv \varepsilon_{zz} \\ &\varepsilon_4 \equiv \varepsilon_{yz}, \quad \varepsilon_5 \equiv \varepsilon_{xz}, \quad \varepsilon_6 \equiv \varepsilon_{xy} \end{aligned} \tag{5.52}$$

There are thus six different numbers present in the strain tensor. The symmetric form of the strain tensor prevents rotation of the unit cell with respect to the Cartesian axis system. It is possible to use this matrix to relate how a vector $\mathbf{r}$ in the unstrained matrix is related to one $\mathbf{r}'$ in the strained structure as follows:

$$\mathbf{r}' = (\mathbf{I} + \boldsymbol{\varepsilon})\mathbf{r} \tag{5.53}$$

$\mathbf{I}$ is the identity matrix. The six first derivatives of the energy with respect to the strain components ε_i measure the forces acting on the unit cell. When combined with the atomic coordinates we get a matrix with $3N + 6$ dimensions. At a minimum not only should there be no force on any of the atoms but the forces on the unit cell should also be zero. Application of a standard iterative minimisation procedure such as the Davidon-Fletcher-Powell method will optimise all these degrees of freedom to give a strain-free final structure. In such procedures a reasonably accurate estimate of the initial inverse Hessian matrix is usually required to ensure that the changes in the atomic positions and in the cell dimensions are matched.

Two common properties which can be calculated from the minimum-energy structure are the elastic and dielectric constants. The elastic constant matrix is used to relate the strains of a material to the internal forces, or stresses It is defined as the second derivative of the energy with respect to the strain, normalised by the cell volume. The inverse of the elastic

constant matrix gives the constant of proportionality between the stress and the strain The elastic constant matrix has dimensions 6×6 and is given by the following expression:

$$\mathbf{C} = \frac{1}{V}[\mathscr{V}''_{\varepsilon\varepsilon} - (\mathscr{V}''_{\varepsilon r} \cdot \mathscr{V}''^{-1}_{rr} \cdot \mathscr{V}''_{r\varepsilon})] \tag{5 54}$$

In this equation $\mathscr{V}''_{\varepsilon\varepsilon}$ is the 6×6 matrix of second derivatives (elements $\partial^2\mathscr{V}/\partial\varepsilon_{ij}^2$), $\mathscr{V}''_{\varepsilon i}$ and $\mathscr{V}''_{r\varepsilon}$ are the corresponding $3N \times 6$ and $6 \times 3N$ mixed coordinate/strain matrices, $\mathscr{V}''_{rr}$ is the $3N \times 3N$ second-derivative coordinate matrix and V is the unit cell volume. It is the second term in Equation (5.54) that accounts for internal atomic relaxations as the cell distorts.

The strains on the lattice are equal to the stress divided by the elastic constant matrix:

$$\boldsymbol{\varepsilon} = (P_{\text{static}} + P_{\text{applied}}) \cdot \mathbf{C}^{-1} \tag{5.55}$$

Here we have expressed the stress as the sum of the (external) applied pressure P_{applied} together with a static pressure P_{static}, which arises from the internal forces acting on the unit cell.

The dielectric constant is concerned with the electrical properties of a material. The dielectric constant for a solid is a 3×3 matrix with different components according to the Cartesian axes. These elements are given by:

$$D_{ij} = \delta_{ij} + \frac{4\pi}{V}\mathbf{q}^{\mathrm{T}} \cdot \mathscr{V}''^{-1}_{rr} \cdot \mathbf{q} \tag{5.56}$$

In Equation (5.56) i and j are one of x, y or z; δ_{ij} is the delta function (i.e. equal to one when $i \equiv j$ and zero otherwise) and $\mathbf{q}$ is a vector containing the charges of each species. It is well known that the effect of a dielectric changes in an oscillating electric field (at high enough frequencies the permanent dipoles in the material are unable to keep up with the rapidly changing field). Thus one usually calculates two sets of dielectric constant matrices, corresponding to the low- and high-frequency regimes. If polarisation is included via a shell model (see Section 4.22.2) then both the cores and the shells are used to determine the low-frequency dielectric matrix; at high frequency only the shells are considered.

Comparison of the relative energies following a minimisation calculation can enable predictions to be made of the likely structure for a given material. In the same way that an organic molecule may be able to exist in more than one three-dimensional structure (or conformation – see Chapter 9) so a solid may (in principle) be able to adopt more than one three-dimensional arrangement of its atoms whilst still maintaining a periodic lattice structure. Silica, SiO_2, has been the subject of considerable attention using these methods. The lowest-energy form is α-SiO_2, or quartz. However, it can also form more open structures. A number of such microporous structures are in principle available, three being silicalite, mordenite and faujasite. In one study the energies of these structures relative to the quartz structure were found to be approximately 2.6, 4.9 and 5.1 kcal/mol, respectively [Ooms *et al.* 1988]. Indeed, the silicalite structure is the only one which can be prepared as the pure silicon oxide; the other forms usually require a high aluminium content and are more traditional zeolites. In an extension of this work two slightly different forms of the silicalite structure were simulated. The normal form at room temperature has orthorhombic

symmetry but at low temperatures this changes to monoclinic. These two forms are very closely related, differing only by the distortion of a key angle by 0.64°. Nevertheless, an energy-minimisation calculation starting from the orthorhombic structure did indeed change to the monoclinic, in agreement with the experimental data [Bell *et al.* 1990]. The orthorhombic → monoclinic transition could only be observed using a force field which included polarisation effects (i.e. the shell model). Lattice minimisation methods can sometimes be very useful in helping to solve the structure of materials, a noteworthy example being the determination of the structure of a zeolite NU-87 [Shannon *et al.* 1991]. This synthetic material is of particular interest as a catalyst as it contains a multidimensional channel system. Multidimensional systems permit more complex catalytic reactions to occur and are also less prone to deactivation than one-dimensional systems. In this case, there are rings containing ten and twelve oxygen atoms (Figure 5.36 (colour plate section)). NU-87 also has a high silica content, which confers improved stability to heat. A number of experimental techniques were used to try to determine the structure, including electron diffraction and powder synchrotron X-ray diffraction, as a result of which an approximate structure was deduced but there remained some features in the powder diffraction spectrum that could not be accounted for These were initially believed to be due to impurities but after energy-minimisation studies some subtle changes in the structure occurred to give a related structure that was a better match to the experimental data A key feature of this particular minimisation was that the structure was not forced to adopt any specific symmetry but rather each atom was able to move independently of the others.

The calculation of vibrational frequencies (called *phonons*) is important to the study of the solid state. Indeed, the calculation of and study of phonons is often given a special name, *lattice dynamics*. To calculate the vibrational frequencies for a solid one follows a very similar approach to that described earlier for molecules, with the exception that when a shell model is being used* then their effect must be incorporated into the mass-weighted matrix of second derivatives (though not directly as they have no mass):

$$\mathbf{W}'' = \mathbf{W}''_{\text{core-core}} - \mathbf{W}''_{\text{core-shell}} \cdot \mathbf{W}''^{-1}_{\text{shell-shell}} \cdot \mathbf{W}''_{\text{core-shell}} \tag{5.57}$$

Of additional importance is that the vibrational modes are dependent upon the reciprocal lattice vector $\mathbf{k}$. As with calculations of the electronic structure of periodic lattices these calculations are usually performed by selecting a suitable set of points from within the Brillouin zone. For periodic solids it is necessary to take this periodicity into account; the effect on the second-derivative matrix is that each element ij needs to be multiplied by the phase factor $\exp(i\mathbf{k} \cdot \mathbf{r}_{ij})$. A *phonon dispersion curve* indicates how the phonon frequencies vary over the Brillouin zone, an example being shown in Figure 5.37. The phonon density of states is the variation in the number of frequencies as a function of frequency. A purely transverse vibration is one where the displacement of the atoms is perpendicular to the direction of motion of the wave; in a purely longitudinal vibration the atomic displacements are in the direction of the wave motion Such motions can be observed in simple systems (e.g. those that contain just one or two atoms per unit cell) but for general three-dimensional lattices most of the vibrations are a mixture of transverse and longitudinal motions, the exceptions

* The use of a shell model is generally recommended otherwise the resulting frequencies are too high.

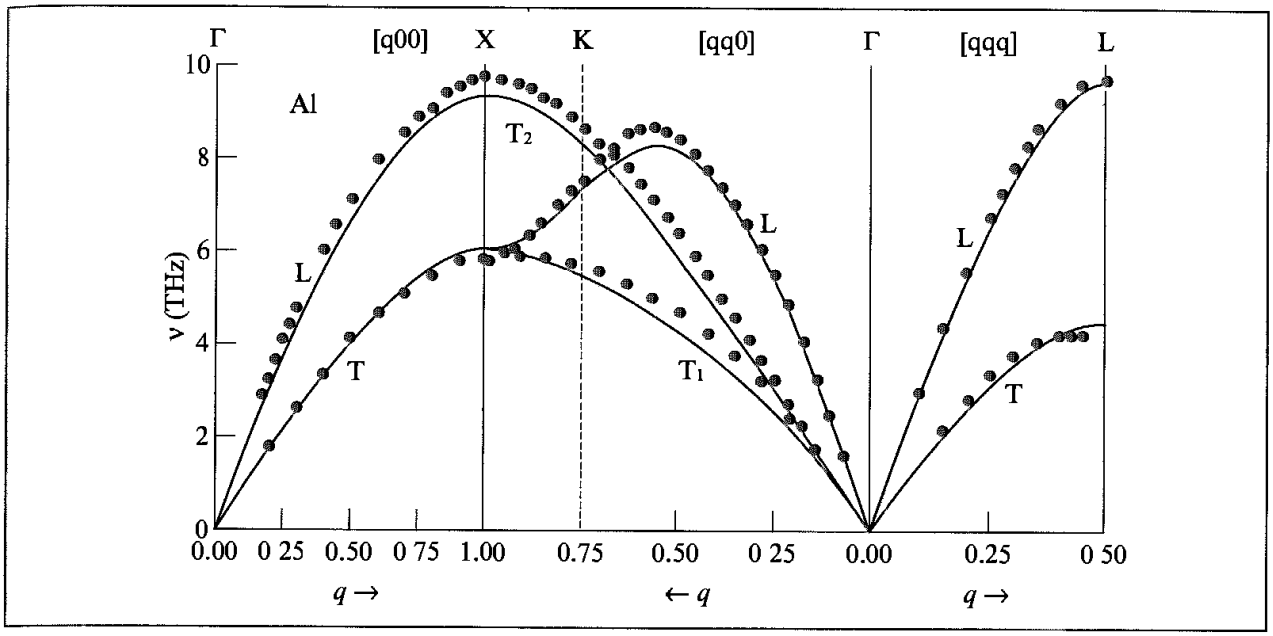

Fig 5 37· Comparison of the calculated phonon dispersion curve for Al with the experimental values measured using neutron diffraction (Figure redrawn from Michin Y, D Farkas, M J Mehl and D A Papaconstantopoulos 1999 Interatomic Potentials for Monomatomic Metals from Experimental Data and ab initio *Calculations* Physical Review ***B59*** *3393–3407)*

being those along directions of high symmetry. The phonons are additionally classified as acoustic or optical; the former are typically of longer wavelength (lower-frequency oscillations) where the atoms move as a unit. The name arises from the fact that these are often measured as sound waves. At the point $\mathbf{k} = 0$ (the gamma point) the first three vibrational frequencies correspond to translation of the entire lattice. The optical phonons are typically higher in frequency. Various experimental techniques can be used to investigate lattice vibrations and to determine the phonon dispersion curves, the most powerful of which is inelastic scattering using thermal neutrons. These often allow the entire range of $\mathbf{k}$ to be sampled, in contrast to some of the alternative types of radiation.

Once the phonon frequencies are known it becomes possible to determine various thermodynamic quantities using statistical mechanics (see Appendix 6.1). Here again some slight modifications are required to the standard formulae. These modifications are usually a consequence of the need to sum over the points sampled in the Brillouin zone. For example, the zero-point energy is:

$$U_{\text{vib}}(0) = \sum_{q=1}^{p} w_q \sum_{i=1}^{N_{\text{nm}}} \frac{h\nu_i}{2} \tag{5.58}$$

In Equation (5.58) the outer summation is over the p points q which are used to sample the Brillouin zone, w_q is the fractional weight associated with each point (related to the volume of Brillouin zone space surrounding q) and ν_i are the phonon frequencies. In addition to the internal energy due to the vibrational modes it is also possible to calculate the vibrational entropy, and hence the free energy. The Helmholtz free energy at a temperature

T is thus given in the quasi-harmonic approximation by the sum of static and vibrational contributions:

$$A = \mathscr{V} + \sum_{q=1}^{p} w_q \sum_{i=1}^{N_{nm}} \left(\frac{h\nu_i}{2} + k_B T \ln \left[1 - \exp\left(-\frac{h\nu_i}{k_B T} \right) \right] \right) \tag{5.59}$$

Here, $\mathscr{V}$ is the internal energy calculated from the potential energy model. The heat capacity at constant volume is another useful thermodynamic quantity that can be determined directly from the frequencies as it equals the derivative of the vibrational internal energy with respect to temperature.

An extension of these ideas involves minimisation of the free energy as a function of the coordinates and the temperature. The function to be minimised is sometimes referred to as an *availability*, given by $G^* = A + P_{ext}V$ where P_{ext} is the external pressure and V is the volume. Such a free energy minimisation requires derivatives of the free energy with respect to the coordinates. Early implementations used approximations such as separating the changes in the external coordinates (i.e. the dimensions of the unit cell) from the internal coordinates (i.e. the locations of the ions within the unit cell). In addition, true free energy derivatives might be calculated for the external coordinates only (due to the computational cost) with the internal coordinates being changed using the static potential energy It is now possible to calculate a full set of analytical free energy first derivatives and hence to perform a full minimisation of the free energy with respect to external and internal coordinates simultaneously [Taylor *et al.* 1998].

Free energy minimisation provides information that is in many ways complementary to molecular dynamics simulations [Allan *et al.* 2000]. The former is particularly useful for investigating materials at lower temperatures where the harmonic assumption is valid; moreover it includes zero-point energy and quantisation effects, which are ignored by molecular dynamics. In addition, free energy minimisation provides free energies directly, rather than free energy differences, and is computationally significantly less expensive. Conversely anharmonic effects become important at higher temperatures, making molecular dynamics and Monte Carlo more suitable. One property that can be calculated via free energy minimisation is the thermal expansivity. This involves a series of free energy minimisations at different temperatures. It is also possible to calculate the free energy of disordered solids, and thus enthalpies and entropies of mixing.

Further Reading

Catlow C R A 1998. Solids: Computer Modelling In Schleyer, P v R, N L Allinger, T Clark, J Gasteiger, P A Kollman, H F Schaefer III and P R Schreiner (Editors) *The Encyclopedia of Computational Chemistry*, John Wiley & Sons, Chichester.

Gill P E and W Murray 1981. *Practical Optimization* London, Academic Press.

McKee M L and M Page 1993. Computing Reaction Pathways on Molecular Potential Energy Surfaces In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 4 New York, VCH Publishers, pp. 35–65

Press W H, B P Flannery, S A Teukolsky and W T Vetterling 1992 *Numerical Recipes in Fortran.* Cambridge, Cambridge University Press.

Schlegel H B 1987. Optimization of Equilibrium Geometries and Transition Structures In Lawley K P (Editor) *Ab Initio Methods in Quantum Chemistry - I* New York, John Wiley & Sons, pp 249-286.

Schlegel H B 1989 Some Practical Suggestions for Optimizing Geometries and Locating Transition States In Bertrán J and I G Csizmadia (Editors). *New Theoretical Concepts for Understanding Organic Reactions* Dordrecht, Kluwer, pp. 33-53

Schlick T 1992 Optimization Methods in Computational Chemistry. In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 3. New York, VCH Publishers, pp 1-71

Stassis C 19 Lattice Dynamics. In Sköld and D L Price (Editors) *Methods of Experimental Physics Volume 23. Neutron Scattering Part A.* Orlando, Academic Press, pp. 369-440.

Watson G W, P Tschaufeser, A Wall, R A Jackson and S C Parker 1997. Lattice Energy and Free Energy Minimisation Techniques. *Computer Modelling in Inorganic Crystallography.* San Diego, Academic Press, pp 55-81.

Williams I H 1993. Interplay of Theory and Experiment in the Determination of Transition-State Structures. *Chemical Society Reviews* **1**:277-283.

References

Allan N L, G D Barrera, J A Purton, C E Sims and M B Taylor 2000. Ionic Solids at High Temperatures and Pressures: *Ab initio*, Lattice Dynamics and Monte Carlo Studies. *Physical Chemistry Chemical Physics* **2**.1099-1111.

Ayala P Y and H B Schlegel 1997 A Combined Method for Determining Reaction Paths, Minima and Transition State Geometries *Journal of Chemical Physics* **107** 375-384

Baker J 1986. An Algorithm for the Location of Transition States. *Journal of Computational Chemistry* 7:385-395.

Bell R G, R A Jackson and C R A Catlow 1990. Computer Simulation of the Monoclinic Distortion in Silicalite. *Journal of the Chemical Society Chemical Communications* **10**.782-783.

Brooks B and M Karplus 1983 Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor. *Proceedings of the National Academy of Sciences USA* **80**:6571-6575

Czerminski R and R Elber 1990. Self-Avoiding Walk Between 2 Fixed-Points as a Tool to Calculate Reaction Paths in Large Molecular Systems. *International Journal of Quantum Chemistry* **S24**:167-186.

Dauber-Osguthorpe P, V A Roberts, D J Osguthorpe, J Wolff, M Genest and A T Hagler 1988 Structure and Energetics of Ligand Binding to Proteins· *Escherichia coli* Dihydrofolate Reductase-Trimethoprim, A Drug-Receptor System. *Proteins Structure, Function and Genetics* 4:31-47

Doubleday C, J McIver, M Page and T Zielinski 1985 Temperature Dependence of the Transition-State Structure for the Disproportionation of Hydrogen Atom with Ethyl Radical. *Journal of the American Chemical Society* **107**:5800-5801

Elber R and M Karplus 1987. A Method for Determining Reaction Paths in Large Molecules: Application to Myoglobin. *Chemical Physics Letters* **139**·375-380.

Fischer S, P D J Groothenuis, L C Groenen, W P van Hoorn, F C J M van Geggel, D N Reinhoudt and M Karplus 1995 Pathways for Conformational Interconversion of Calix[4]arenes. *Journal of the American Chemical Society* **117**:1611-1620.

Fischer S and M Karplus 1992 Conjugate Peak Refinement· An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. *Chemical Physics Letters* **194**:252-261.

Fisher C L, V A Roberts and A T Hagler 1991. Influence of Environment on the Antifolate Drug Trimethoprim: Energy Minimization Studies *Biochemistry* **30**:3518–3526

Fukui K 1971. Recognition of Stereochemical Paths by Orbital Interaction. *Accounts of Chemical Research* 4.57–64

Fukui K 1981. The Path of Chemical Reactions – The IRC Approach *Accounts of Chemical Research* **14**:368–375.

Gelin B R and M Karplus 1975. Sidechain Torsional Potential and Motion of Amino Acids in Proteins: Bovine Pancreatic Trypsin Inhibitor *Proceedings of the National Academy of Sciences USA* **72** 2002–2006.

Gonzalez C and H B Schlegel 1988. An Improved Algorithm for Reaction Path Following *Journal of Chemical Physics* **90**.2154–2161.

Houk K N, J González and Y Li 1995 Pericyclic Reaction Transition States: Passions and Punctilios 1935–1995 *Accounts of Chemical Research* **28**·81–90

Houk K N, Y Li and J D Evanseck 1992 Transition Structures of Hydrocarbon Pericyclic Reactions. *Angewandte Chemie International Edition in English* **31**·682–708.

Nowak W, R Czerminski and R Elber 1991. Reaction Path Study of Ligand Diffusion in Proteins: Application of the Self Penalty Walk (SPW) Method to Calculate Reaction Coordinates for the Motion of CO through Leghemoglobin *Journal of the American Chemical Society* **113** 5627–5737.

Ooms G, R A van Santen, C J J Den Ouden, R A Jackson and C R A Catlow 1988. Relative Stabilities of Zeolitic Aluminosilicates. *Journal of Physical Chemistry* **92** 4462–4465.

Peng C, P Y Ayala, H B Schlegel and M J Frisch 1996. Using Redundant Internal Coordinates to Optimise Equilibrium Geometries and Transition States. *Journal of Computational Chemistry* **17**:49–56.

Pople J A, A P Scott, M W Wong and L Radom 1993. Scaling Factors for Obtaining Fundamental Vibrational Frequencies and Zero-Point Energies from HF/6-31G* and MP2/6-31G* Harmonic Frequencies *Israel Journal of Chemistry* **33**:345–350.

Schlegel H B 1982. Optimisation of Equilibrium Geometries and Transition Structures *Journal of Computational Chemistry* **3**·214–218

Shannon M D, J L Casci, P A Cox and S J Andrews 1991. Structure of the Two-dimensional Medium-pore High-silica Zeolite NU-87. *Nature* **353**:417–420

Taylor M B, G D Barrera, N L Allan, T H K Barron and W C Mackrodt 1998. Shell: A Code for Lattice Dynamics and Structure Optimisation of Ionic Crystals. *Computer Physics Communications* **109**: 135–143.

Woodward R B and R Hoffmann 1969. The Conservation of Orbital Symmetry. *Angewandte Chemie International Edition in English* **8**:781–853.